SLOVAK UNIVERSITY OF TECHNOLOGY IN BRATISLAVA

Faculty of Electrical Engineering and Information Technology

Reg. No.: FEI-104354-82208

Probabilistic modelling and functional data analysis of sleep structure

Dissertation thesis

2018 **Mgr. Zuzana Rošťáková**

SLOVAK UNIVERSITY OF TECHNOLOGY IN BRATISLAVA

Faculty of Electrical Engineering and Information Technology

Reg. No.: FEI-104354-82208

Probabilistic modelling and functional data analysis of sleep structure

Dissertation thesis

Study programme: Measurement Technology
Study field: 5.2.54. Measurement Technique
Training workplace: Institute of Measurement Science of the Slovak Academy of Science
Supervisor: Ing. Mgr. Roman Rosipal, PhD.

Bratislava 2018 **Mgr. Zuzana Rošťáková**

Slovak University of Technology in Bratislava
Institute of Electrical Engineering

Faculty of Electrical Engineering and Information Technology
Academic year: 2017/2018
Reg. No.: FEI-104354-82208

STU FEI

DISSERTATION THESIS TOPIC

Student:	**Mgr. Zuzana Rošťáková**
Student's ID:	82208
Study programme:	Measurement Technology
Study field:	5.2.54. Measurement Technique
Thesis supervisor:	Ing. Mgr. Roman Rosipal, PhD.
External educational institution:	Institute of Measurement Science of the Slovak Academy of Science

Topic: **Probabilistic modelling and functional data analysis of sleep structure**

Language of thesis: English

Specification of Assignment:

Sleep deprivation, whether from disorder or lifestyle, whether acute or chronic, poses a significant risk in day-time cognitive performance, excessive somnolence, fatigue, or impaired vigilance. The aim of the dissertation is to use new probability models of the sleep process, functional data analysis and machine learning methodologies to objectify the quality of sleep. An important element of this goal is the definition and search of significant sleep biomarkers correlated with the daily cognitive, physiological and psychological status and performance of observed subjects. In addition to a healthy population, it is intended to study patients with specific brain lesions as it is known that such patients are susceptible to sleep disorders that often lead to disorders in daytime performance and attention.

Assignment procedure from:	25. 04. 2016
Date of thesis submission:	31. 05. 2018

Mgr. Zuzana Rošťáková
Solver

prof. Ing. Viktor Smieško, PhD.
Head of department

prof. Ing. Robert Redhammer, PhD.
Study programme supervisor

Acknowledgement

First, I would like to thank my home supervisor Dr. Roman Rosipal for the continuous support of my PhD study. His ideas, recommendations and remarks were very helpful all the time of the research. Special thanks for patience control and finalisation of my articles and dissertation thesis.

I would like to also thank ao. Univ. Prof. Dr. Georg Dorffner for the opportunity to join his team at the Section for Artificial Intelligence and Decision Support, Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna and his remarks on the interpretation of the time alignment of the sleep probabilistic curves.

Many thanks go to my colleagues for their friendly communication, motivational words and willingness to help.

Last but not least, I thank my family. Their continuous support and tolerance encouraged me to progress all the time of my study.

This research was supported by the Slovak Research and Development Agency (grant number APVV-16-0202), the Ministry of Health of the Slovak Republic (grant number MZ 2013/46-SAV-6) and by the Scientific Grant Agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic and Slovak Academy of Sciences (VEGA, grant number 2/0011/16).

Abstrakt

ROŠŤÁKOVÁ, Zuzana: Pravdepodobnostné modelovanie a funkcionálna dátová analýza štruktúry spánku. [Dizertačná práca] – Slovenská technická univerzita v Bratislave. Fakulta elektrotechniky a informatiky.
Školiteľ: Ing. Mgr. Roman Rosipal, Ph.D.
Bratislava 2018, 128 strán.

Spánok je spojitý heterogénny proces, ktorý počas noci prechádza konečným počtom spánkových stavov. Jeho kvalita a štruktúra vo výraznej miere ovplyvňujú naše každodenné správanie. Spánkový elektroencefalogram (EEG) zachytávajúci aktivitu mozgu počas spánku tvorí základný kameň tejto práce. Nameraný EEG signál je následne spracovaný pomocou pravdepodobnostného spánkového modelu a reprezentovaný konečnou množinou spánkových pravdepodobnostných kriviek. Prvá časť práce je zameraná na detekciu spánkových profilov, ktoré významne súvisia s dennými mierami (subjektívne hodnotenie kvality spánku, fyziologický stav organizmu, kognitívne testovanie) pomocou metód funkcionálnej dátovej analýzy, konkrétne zhlukovej analýzy spánkových kriviek. Ak spánkové krivky nie sú synchronizované v čase, zhluková analýza môže viesť k zaradeniu kriviek s podobným profilom do rôznych zhlukov. Existujúce metódy simultánne kombinujúce zhlukovanie a synchronizáciu kriviek pri aplikácii na spánkové dáta nevedú k uspokojivým výsledkom. Z tohto dôvodu sme navrhli vlastnú metódu, ktorá iteračne kombinuje zhlukovú analýzu kriviek a časovú synchronizáciu. Jej benefity oproti existujúcim prístupom sú demonštrované na dvoch množinách spánkových dát. Vzniknuté zhluky obsahujú dva typy variability – medzi zhlukmi a v rámci zhlukov. Na detekciu a analýzu oboch druhov variability sme si zvolili viacstupňovú funkcionálnu verziu metódy hlavných komponentov. Táto metóda bola pôvodne určená len pre dáta s rovnakým počtom pozorovaní v rámci zhlukov. V tejto práci uvádzame aj jej verziu rozšírenú na prípad, keď sa počet pozorovaní medzi zhlukmi líši.

Kľúčové slová: spánkový elektroencefalogram, funkcionálna dátová analýza, pravdepodobnostný spánkový model, synchronizácia kriviek, zhluková analýza kriviek, viacstupňová funkcionálna verzia metódy hlavných komponentov.

Abstract

ROŠŤÁKOVÁ, Zuzana: Probabilistic modelling and functional data analysis of sleep structure. [Dissertation thesis] – Slovak University of Technology in Bratislava. Faculty of Electrical Engineering and Information Technology.
Supervisor: Ing. Mgr. Roman Rosipal, Ph.D.
Bratislava 2018, 128 pages.

Sleep is a continuous heterogeneous process consisting of a finite number of sleep stages during a night. Its quality and structure influences our daily behaviour. The sleep electroencephalogram (EEG) representing electrical activity of a brain during sleep is a cornerstone of the thesis. The EEG signal is processed by the probabilistic sleep model and represented by a finite set of sleep probabilistic curves. In the first part of the thesis we aim to detect specific sleep profiles – sleep biomarkers – which significantly correlate with measures representing daily behaviour (subjectively scored sleep quality, physiological factors, cognitive tests). For this purpose functional data analysis – namely the curve cluster analysis – is used. However, when the sleep curves are misaligned in time, cluster analysis may assign curves with a similar profile into different clusters. Existing methods which simultaneously combine cluster analysis and curves alignment are not appropriate for our sleep data. Therefore we propose a novel method based on iterative combination of curves alignment and clustering. Benefits of our method over existing approaches are demonstrated on two datasets of sleep curves. Formed clusters show two types of variability – between and within clusters. To detect and analyse these types of variability the multilevel functional principal component analysis is used. This method was originally developed only for data with the same number of observations within clusters. In this thesis we propose its modification for data with varying number of observations per cluster.

Key words: sleep electroencephalogram, functional data analysis, probabilistic sleep model, curve registration, curve clustering, multilevel functional principal component analysis.

Contents

List of abbreviations

Sleep monitoring

A1, A2	reference electrodes located on the ear lobes
C3, C4	electrodes located over scalp in the central region
ECG	electrocardiogram
EEG	electroencephalogram
EMG	electromyogram
EOG	electrooculogram
Fp1, Fp2	electrodes located over scalp in the frontal region
M1, M2	reference electrodes located on the *mastoid*
O1, O2	electrodes located over scalp in the occipital region
PSG	polysomnography
SIESTA	the SIESTA project polygraphic and clinical database [Klösch et al., 2001]

Daily measures for healthy sleepers

SRQ_aq	Self–rating questionnaire for awakening quality [Saletu et al., 1987]
SRQ_scom	Self–rating questionnaire for somatic complaints
SRQ_sq	Self–rating questionnaire for sleep quality
VAS_aff,	Visual analogue scale test for affectivity [Aitken, 1969]
VAS_drive,	Visual analogue scale test for drive
VAS_drows	Visual analogue scale test for drowsiness
VAS_mood	Visual analogue scale test for mood
WB_m, WB_e	Well–being self assessment scale (morning, evening) [von Zerssen et al., 1970]
dia_m, dia_e	diastolic blood pressure (morning, evening)

pul_m, pul_e	pulse rate (morning, evening)
sys_m, sys_e	systolic blood pressure (morning, evening)
ACT_ts	Alphabetical cross-out test, total score [Grünberger, 1977]
ACT_sv	Alphabetical cross-out test, attention variability
ACT_errp	Alphabetical cross-out test, % of errors
FMAT_r	Fine motor activity test (right hand) [Grünberger, 1977]
FMAT_l	Fine motor activity test (left hand)
NMT	Numerical memory test [Grünberger, 1977]

Daily measures for patients after ischemic stroke

FMAT2_cp$n$	Fine motor activity test, number of correctly retraced pixels in pattern n
FMAT2_sp$n$	Fine motor activity test, number of successes in pattern n
LANT_A	Lateralised attention network test, alerting [Greene et al., 2008]
LANT_C	Lateralised attention network test, conflict
LANT_O	Lateralised attention network test, orienting
LANT_OF	Lateralised attention network test, orienting facilitatory
LANT_OI	Lateralised attention network test, orienting inhibitory
LANT_RVF_*type*	Lateralised attention network test, right visual field + *type* (A,C,O,OF,OI)
LANT_LVF_*type*	Lateralised attention network test, left visual field + *type* (A,C,O,OF,OI)
NIHSS	National Institutes of Health Stroke Scale [Brott et al., 1989]
RTT_n	Reaction time test, average reaction time in a trial n
RTT_min	Reaction time test, the minimal reaction time
RTT_mean	Reaction time test, the mean reaction time across trials
T-MENSTAT_A_n	T-MENSTAT questionnaire about subjective feelings after awakening, performed before neurocognitive testing, part n
T-MENSTAT_B_n	T-MENSTAT questionnaire, performed after neurocognitive testing, part n
WMT_bw	Working memory test, backward part [Kaufman and Lichtenberger, 2005]
WMT_bw	Working memory test, forward part

Sleep modelling

a	vector of autoregressive coefficients
AR(n)	autoregressive model of order n
AASM	American Academy of Sleep Medicine
PSM	Probabilistic sleep model
R&K	Rechtschaffen and Kales sleep model
$\mathcal{N}(.\|\mu, \Sigma)$	probability distribution function of a normal distribution with mean μ and covariance matrix Σ
NREM, nonREM	sleep stages without eye movements, *S1, S2* and *SWS*
REM	sleep stage, characteristics with rapid eye movements
S1	sleep stage, light sleep
S2	sleep stage, characteristics with spindles
SWS	slow wave sleep, sleep stage representing deep sleep
Wake	state of full awakening during the night

Functional data clustering

$\mathcal{F}$	functional space
FDA	functional data analysis
d_2	L^2 distance for curves
sil	silhouette
t	time
T	time interval
X	a curve or functional datum defined over a time interval T
$\varphi_1, \ldots, \varphi_L$	basis functions of the functional space $\mathcal{F}$

Time synchronisation of curves

$B_1, \ldots, B_L$	set of L B–splines [Wahba, 1990]
$h : T \to T$	warping function
$\mathcal{H}$	set of all strictly increasing transformations of the time interval T (warping functions)
ETW	Elastic time warping [Tucker et al., 2013]
PCS	Pairwise curve synchronisation [Müller and Tang, 2008]
SMTW	Self–modelling time warping [Gervini and Gasser, 2004]
srsf	square–root slope function

Methods for combination of curves alignment and cluster analysis

DTW	Dynamic time warping
JPCCA	Joint probabilistic curve clustering and alignment [Gaffney and Smyth, 2005]
KMACC	k–mean alignment for curve clustering [Sangalli et al., 2010]
tPCS	truncated version of the Pairwise curve synchronisation method [Tang and Müller, 2009]
2DTW–ETW	2–step approach with the Elastic time warping method used in the registration step
2DTW–PCS	2–step approach with the Pairwise curve synchronisation used in the registration step
2DTW–SMTW	2–step approach with the Self–modelling time warping method used in the registration step

Multilevel functional principal component analysis

FPCA	Functional principal component analysis
MFPCA	Multilevel functional principal component analysis [Di et al., 2009]
I	number of subjects included in the analysis
J, J_i	number of observations for the i^{th} subject
X_{ij}	the j^{th} observation for the i^{th} subject
μ	the overall mean function
η_j	the visit–specific deviation from the overall mean function μ
Z_i	the subject–specific deviation from the visit–specific profile of the i^{th} subject
W_{ij}	the residual deviation from the subject– and visit–specific profile
R_T	the overall covariance function
R_W	the covariance function of difference between two observations within the same subject
R_Z	the covariance function between two observations within the same subject
$\phi_1^{(1)}, \ldots, \phi_{P_1}^{(1)}$	first P_1 functional principal components on level 1
$\phi_1^{(2)}, \ldots, \phi_{P_2}^{(2)}$	first P_2 functional principal components on level 2
$\alpha_{i1}, \ldots, \alpha_{iP_1}$	principal component scores on level 1 for the i^{th} subject

$\beta_{ij1}, \ldots, \beta_{ijP_2}$	principal component scores on level 2 for the i^{th} subject and the j^{th} visit
$\lambda_1^{(1)}, \ldots, \lambda_{P_1}^{(1)}$	the level 1 eigenvalues, eigenvalues of R_Z
$\lambda_1^{(2)}, \ldots, \lambda_{P_2}^{(2)}$	the level 2 eigenvalues, eigenvalues of R_W

Introduction

Sleep is a dynamical process which plays important role in our lives. Its structure, quality and length influences humans daily behaviour, affectivity, mood and also health. The main goal of our long term research is to detect specific sleep profiles – sleep biomarkes – reflecting the sleep quality and significantly correlating with daily life performance.

In order to analyse the sleep process we need an appropriate sleep model. Nowadays, the Rechtschaffen and Kales sleep model [Rechtschaffen and Kales, 1968] or its novel version produced by the American Academy of Sleep Medicine (AASM) [Iber et al., 2007] are widely used in clinical practice. Both models are based on assignment of 30–second long time intervals of sleep into sleep stages according to changes in structure and depth of sleep. However, this discrete representation of sleep shows several disadvantages [Himanen and Hasan, 2000]. Small set of sleep stages, non–smooth changes between sleep stages or too long time segments considered by both models represent only a few of such disadvantages.

Therefore in this thesis we prefer to use the Probabilistic sleep model (PSM) [Lewandowski et al., 2012] which provides a continuous description of the sleep process. The PSM overcomes all above mentioned disadvantages of the standard sleep models because it operates only on three–second long time segments for each of which probability values of their relationship to one of a higher number of sleep states – called microstates – is computed. Considering the probability values of a sleep microstate as a function of time we obtain a sleep probabilistic curve.

Current studies dealing with relationship between sleep structure and humans' physiological state or well-being measures are based on the extraction of one–dimensional sleep characteristics (for example the total sleep time, time spent in each sleep stage separately or sleep efficiency) from the standard sleep models and their correlation with variables representing daily life behaviour. The same procedure is possible also in the case of the PSM – the sleep probabilistic curves can be represented by a set of one–dimensional measures, but this may result into the partial loss of important information about the sleep dynamics. Fortunately, character of sleep probabilistic curves offers a way for inspection

of the sleep structure through functional data analysis.

We divide our thesis into two main parts.

In the first part we aim to detect relationship between sleep structure and daily measures representing subjective feelings in the morning, physiological state of an organism and performance of subjects in neuropsychological tests. Out of many techniques available for functional data and their relation to one–dimensional variables or vectors in this thesis we focus on functional cluster analysis. The main idea is to divide a given dataset of subjects into subgroups according to similarity of their sleep probabilistic curves and then to test whether the clusters significantly differ in a given daily measure. Due to a nontrivial character of our sleep data another approaches, for example the functional regression model [Ramsay and Silverman, 2005], don't lead to easily interpretable results.

Many clustering techniques well–known in the multidimensional data analysis can be after small modifications adapted for functional data clustering, for example k–means or k–medoids. Unfortunately, in practice observed curves showing similar profile are often misaligned time. We say, that two curves are misaligned if they have similar shape but their important characteristics like local optima or zero–crossings occur at different time. Therefore clustering methods may consider these curves as dissimilar and assign them into different clusters.

To avoid misinterpretation of results caused by curve misalignment, Ramsay and Silverman [2005] recommend to align or in other words to register or to synchronise data in time immediately before analysis. Because of nontrivial character of the problem, usually depending on a given data type, there are many methods developed for the curve alignment problem. They differ in the definition of a curve similarity and an algorithm for the curve alignment. In this thesis we describe several chosen registration methods for which we observed a good performance on our sleep data.

However, when the dataset consists of curves with many different profiles, registration methods usually lead to two opposite extrema – either they are not able to align the curves properly or the alignment is visually ideal, but at the cost of rapid distortions of the curves profiles. Therefore it would be appropriate to align only curves with a similar shape. However, this in turn leads us back to the cluster analysis of misaligned curves.

Several methods can be found in literature which produce simultaneous alignment and clustering of curves. Unfortunately, we observed that they are not appropriate for our sleep data due to a high complexity of their algorithms or character of the alignment. Therefore we propose a novel algorithm which iteratively combines cluster analysis with

curves alignment. Its superior performance over existing methods is demonstrated on two sleep datasets.

In the second part of the thesis we deal with the Multilevel functional principal component analysis [Di et al., 2009] as a dimensionality reduction technique for functional data with repeated measurements – hierarchical functional data. The original method was proposed only for balanced data when for each object the same number of measurements is available. We adapt the methodology for the case when the number of observations differs among objects (unbalanced data). This adaptation is one of the main contributions of the thesis. Finally, the both versions of the algorithm are applied to a sleep dataset.

To sum up the focus of the dissertation thesis, our main objectives are as follows:

- Analysis of sleep structure provided by the Probabilistic sleep model and methods of functional data analysis in i) dataset of healthy subjects without serious sleep problems and ii) patients following an ischemic stroke with the aim to detect relationships between sleep structure and daily measures (questionnaires about sleep and awakening quality, mood or drowsiness; neurocognitive tests for short–term memory, fine motor activity; pulse rate, blood pressure).

- Adaptation of existing methods or proposition of a new method for simultaneous alignment and cluster analysis of sleep probabilistic curves. Analysis of advantages and disadvantages of the curves alignment for each sleep stage and each sleep microstate separately.

- Analysis of relationships between differences in sleep structure of two nights of subjects and corresponding difference in daily measures by the Multilevel functional principal component analysis (MFPCA).

- Adaptation of the MFPCA method to functional datasets with repeated observations, where i) the order of observations within subjects is exchangeable, and ii) number of observations varies with subjects.

Our thesis is organised as follows: In the first chapter the brief description of polysomnography as a standard tool for sleep monitoring is given. Then, in this thesis used datasets of healthy subjects without serious sleep problems as well as patients following an ischemic stroke are described. Chapter 2 deals with the standard Rechtschaffen and Kales [Rechtschaffen and Kales, 1968] sleep model and the Probabilistic sleep model [Lewandowski et al., 2012] as well as their application to the studied sleep datasets.

Short introduction to functional cluster analysis and several examples of distance measures between curves is given in Chapter 3. Chapters 4 and 5 deal with the curves misalignment problem and existing methods for solving it. The both chapters form the important theoretical part of the thesis. Moreover, Chapter 5 provides detailed description of our novel methodology for iterative combination of curve clustering and alignment. The proposed method is validated and compared with existing approaches in Chapter 6. We also aim to find the answer to the question for which sleep states the curves alignment shows benefits when detecting a relationship with daily measures and for which it is counter–productive.

Next, Chapter 7 is focused on the Multilevel functional principal component analysis [Di et al., 2009] and its adaptation to the both balanced and unbalanced data. Main results and contributions of the thesis are summarised in Conclusion.

Chapter 1

Sleep monitoring

Sleep is a dynamic process which can be described by a finite set of sleep states. In this chapter the polysomnography (PSG) is described as a standard diagnostic tool for the sleep monitoring. In the centre of our interest is a set of electrical biosignals representing the activity of an human organism during sleep, especially the *electroencephalogram* and the *electromyogram*. We focus on their basic properties, measurement and their relevance for the analysis of the sleep process.

In the second part of the chapter two sleep datasets, which are used in this thesis, are described in details – dataset of patients following an ischemic stroke and a control group of healthy sleepers.

1.1 Polysomnography

Polysomnography is a standard diagnostic tool for the sleep study. It includes a set of recordings of the biophysiological changes in the human body occurring during the sleep process. The information about the electrical potentials produced by brain – *electroencephalogram* (EEG) – and the muscle activity – *electromyogram* (EMG) – are the most important parts of the PSG recordings for our thesis. Moreover, PSG includes also the measurement of the eye movements (*electrooculogram*, EOG); the heart rhythm (*electrocardiogram*, ECG) and breathing (*Snore, Flow*) are also monitored.

Recording of the EEG signal is based on the changes of electrical potentials over time between a pair of electrodes located on the scalp – a single electrode and a reference electrode [Bronzino, 2000]. Electrodes usually consist of Ag–AgCl discs with 1–3 mm in diameter. Stability and low impedance in the electrode–skin interface are achieved by

fixing the electrodes with a special contact gel. For a long–term measurement spindle electrodes located between the skin and the skull of a subject are preferred [Bronzino, 2000].

Figure 1.1 represents an example of one minute long EEG signal recording of a healthy subject without serious sleep problems. Information incorporated in the EEG signal may reveal many pathologies in the brain activity, epilepsy or other diseases. Moreover, the EEG signal is useful in the process of the sleep modelling and assignment of short time intervals into sleep stages. For more information see Section 2.1.

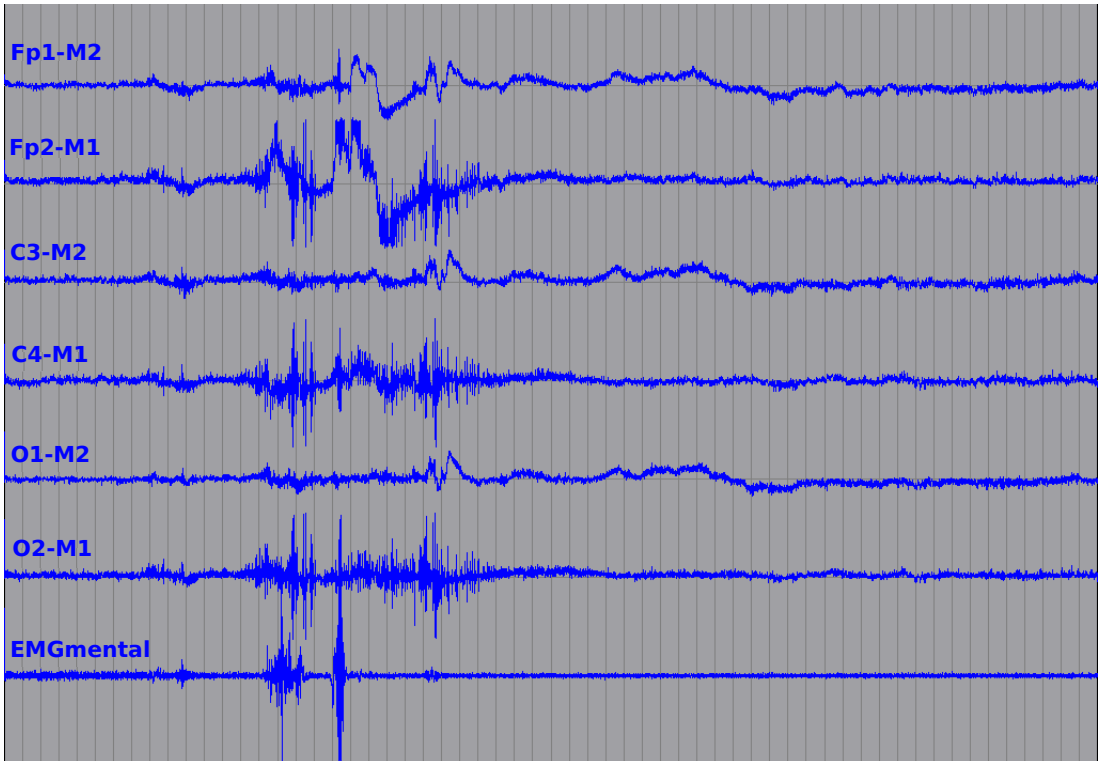


Figure 1.1: One minute long recording of the EEG and EMG signals of a healthy subject without serious sleep problems. In the case of the EEG signal three pairs of electrodes were used – frontal Fp1–M2, Fp2–M1, central C3–M2, C4–M1 and occipital O1–M2, O2–M1. The EMG signal was recorded by two electrodes located above and one located below the subject's chin.

The EMG represents electrical signals travelling between muscles and the nervous system. It is useful for diagnosis of the neuromuscular diseases. In the sleep study it forms a useful tool for identification of the sleep stages which are connected with increased movements. Recording of the EMG signal is obtained by the intramuscular or surface electrodes. The first type of electrodes consists of a thin needle applied to the belly of the

muscle. For the sleep PSG measurement the surface electrodes are preferred. In our thesis the EMG signal was measured by a pair of electrodes located above and one electrode located below the subjects' chin.

1.2 Sleep data

In this section two sleep datasets used in the thesis are described in details – the database of subjects without serious sleep problems and the database of patients after ischemic stroke.

1.2.1 SIESTA database

The European sleep database SIESTA [Klösch et al., 2001] is a systematic PSG database with sleep recordings of 200 healthy subjects and of 100 subjects with selected sleep disorders of high prevalence. In this thesis we used a subset of the SIESTA database including PSG recordings of 146 subjects without serious sleep problems spending two consecutive nights in the sleep lab. The database consists of 85 men and 61 women with the average age of 53 years.

The PSG measurement started right after going to bed and switching the lights off, the recording stopped after a subject awoke spontaneously. The EEG signal was measured by three pairs of electrodes – frontal (Fp1–M2, Fp2–M1), central (C3–M2, C4–M1) and occipital (O1–M2, O2–M1). The location of the electrodes on the scalp is depicted in Figure 1.2. The reference electrodes M1 and M2 were placed on the *mastoid*. Two electrodes for monitoring the muscle activity were placed above the chin of a subject, the reference electrode was applied below the chin. Linked electrodes applied to the *musculus anterior tibialis* on both legs were helpful for another EMG recording.

After awakening subjects filled out several questionnaires scoring their sleep and awakening quality. Subjective sleep and awakening quality was assessed in the morning utilising a standardised Self–rating questionnaire (SRQ) [Saletu et al., 1987]. The SRQ consists of 20 items yielding in three sub–scores (sleep quality, awakening quality and somatic complaints). Four 100 mm visual analogue scales [Aitken, 1969] for drive, mood, affectivity and drowsiness were also used. The self–assessment questionnaire of well–being consisting of 28 items [von Zerssen et al., 1970] was filled by subjects in the evening and morning sessions.

Moreover, the subjects performed several neuropsychologic tests for assessment of attention, attention variability, concentration, short–term memory and fine motor activity [Grünberger, 1977]. Tests were carried out after washing, getting dressed and breakfast and in general between one and two hours after getting up. For more details see [Klösch et al., 2001; Rosipal et al., 2013]. Finally, the evening blood pressure and pulse values were recorded in less than two hours before bedtime and in the morning after sleep. The list of all daily measures and their abbreviations can be found in Table 1.1.

In contrast to [Rosipal et al., 2013] the daily measures listed in Table 1.1 were rescaled in a way that low values indicate good sleep. Several neuropsychological tests and physiological measures showed high correlation with age. Therefore the age effect was adjusted by subtracting an n^{th} order polynomial ($n = 1, 2, 3$) fitted to the data in the least squares sense.

Moreover, following the research lines presented in [Rosipal et al., 2013], we considered three factor scores – factor of subjectively scored sleep and awakening quality (*FA1*), physiological factor (*FA2*) and neuropsychological factor (*FA3*) obtained as the dominant factors after applying factor analysis to the set of daily measures listed in Table 1.1.

1.2.2 Patients following ischemic stroke

In our thesis the database of PSG recordings of patients after ischemic stroke hospitalised at the 1^{st} Department of the Neurology, Comenius University in Bratislava was used. The standard overnight PSG recording took place one to 10 days after a stroke occurred. The recording itself was carried out by the Alice® 5 or Alice® 6 device [Philips Respironics, 2010].

The EEG signal of patients after stroke was measured only by two pairs of electrodes – central (C3–A2, C4–A1) and occipital (O1–A2, O2–A1). The reference electrodes A1 and A2 were placed on the ear lobes (Figure 1.2). The EMG signal was measured in the same way as in the case of the SIESTA database.

The first part of the database of patients after stroke consists of PSG recordings of 60 men (average age 65 years) and 35 women (average age 68 years) measured between years 2013 and 2017. This dataset or its subsets served for the detection of changes in sleep patterns between patients after stroke and healthy sleepers of the same age and gender [Rohleder, 2016; Rošťáková, 2015; Bui, 2014; Bui and Rosipal, 2014; Škoviera et al., 2014] and between patients with different type, location or severity of the stroke [Rohleder, 2016], but it is not considered in this thesis.

Daily measure	Abbreviation
Self–rating questionnaire for sleep quality	SRQ_sq
Self–rating questionnaire for awakening quality	SRQ_aq
Self–rating questionnaire for somatic complaints	SRQ_scom
Visual analogue scale test for drive	VAS_drive
Visual analogue scale test for mood	VAS_mood
Visual analogue scale test for affectivity	VAS_aff
Visual analogue scale test for drowsiness	VAS_drows
Well–being self assessment scale (morning/evening)	WB_m, WB_e
pulse rate (morning/evening)	pul_m, pul_e
systolic blood pressure (morning/evening)	sys_m, sys_e
diastolic blood pressure (morning/evening)	dia_m, dia_e
Numerical memory test	NMT
Alphabetical cross–out test (total score)	ACT_ts
Alphabetical cross–out test (variability)	ACT_sv
Alphabetical cross–out test (% of errors)	ACT_errp
Fine motor activity test (right/left hand)	FMAT_r, FMAT_l

Table 1.1: The list of daily measures of the healthy sleepers from the SIESTA database and their abbreviations.

The second part of the dataset of patients after stroke includes PSG recordings of 24 patients. These patients also took part in a battery of cognitive tests and one questionnaire performed in the morning after the PSG recording. These were

- the Fine motor activity test (FMAT2) where the subjects were asked to redraw the template patterns of simple geometric shapes (Figure 1.3). Then the percentage of correctly retraced pixels was computed. Here we use the abbreviation FMAT2 in order to distinguish between this test and the Fine motor activity test for healthy subjects (FMAT).

- the Lateralised attention network test (LANT) [Greene et al., 2008] measures three independent behavioural components of attention (Table 1.2)

  – spatial Orienting - consists of two parts i) the benefit or facilitatory component of a valid spatial cue (Orienting facilitatory, LANT_OF) and ii) the cost or inhibitory component of an invalid spatial cue (Orienting inhibitory, LANT_OI),

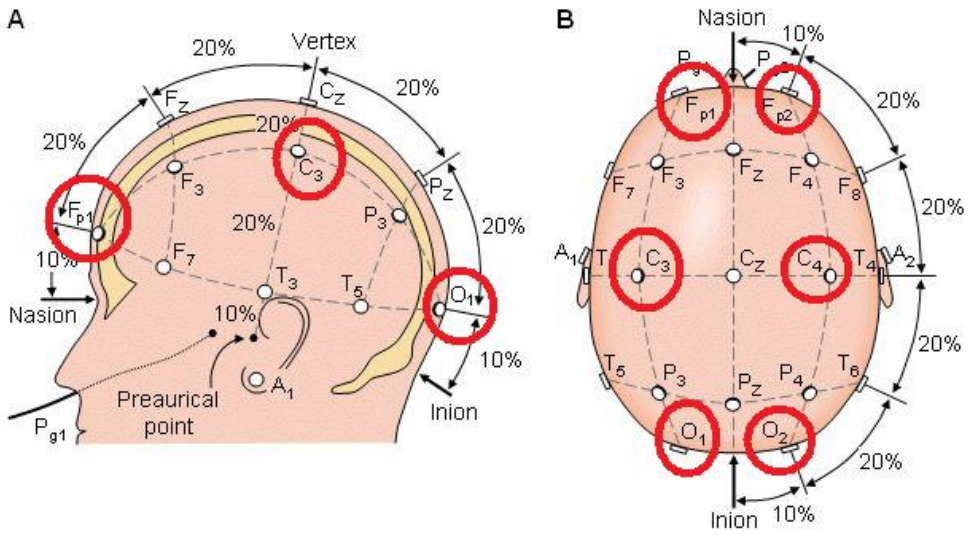


Figure 1.2: The scalp placement of three pairs of electrodes used for the EEG signal recording. Original image can be found in [Malmivuo and Plonsey, 1995].

- Alerting - the benefit of temporal pre–cues (LANT_A),
- Conflict resolution - ability to overcome distracting stimuli (LANT_C).

The test consists of five arrows briefly flashed in the left/right visual field. The middle arrow is a target and the four surrounding arrows form a distractor (congruently or incongruently orienting to the target), the presentation side (right/left) is indicated with a pre-cue $\star$ (valid, invalid, no cue, double cue). The goal of the test is to determine the orientation of the target arrow. Score of each independent component of the test is computed in the following way

- spatial Orienting - difference in reaction times to valid cue (LANT_OF) or invalid cue (LANT_OI) and double cue,
- Alerting - difference in reaction times to double cue and no cue,
- Conflict resolution - difference in reaction times to congruent stimuli and incongruent stimuli.

For more details see [Greene et al., 2008] or [Rybár et al., 2016].

- the Reaction time test (RTT) formed a training exercise before the LANT. A subject was asked to click with a computer mouse once a target (circle) occurred on a computer screen as quickly as possible. According to the subject's use of index

finger or middle finger on the dominant and non–dominant hand we distinguished four trials. The outcome of the test consists of the average reaction time for each trial, mean and minimal reaction time across trials.

- the T–MENSTAT questionnaire [Pacific Development and Technology, LCC, 2012] was filled by subjects immediately before (T–MENSTAT_A) and after performing the neurocognitive tests (T–MENSTAT_B). It consists of four parts describing subjective level of energy and motivation (T–MENSTAT_A/B_1), fatigue (T–MENSTAT_A/B_2), frustration (T–MENSTAT_A/B_3) and drowsiness (T–MENSTAT_A/B_4).

- the Working memory test (WMT) which represents a standard test for assessing the capacity of short–time and working memory [Kaufman and Lichtenberger, 2005]. The subject is asked to repeat a sequence of presented digits in the same or reverse order. Once the correct answer is obtained, the length of a sequence increases. The total score reflects the length of the longest sequence which the subject was able to repeat correctly.

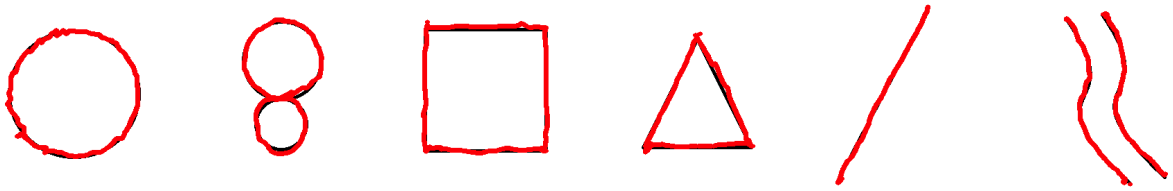


Figure 1.3: Six basic shapes used in the Fine motor activity test (black lines). The shapes were retraced by 68–year–old right–handed man after stroke (red lines).

Severity of a stroke was ranked one, 7 and 90 days after the stroke occurrence according to the National Institutes of Health Strokes Scale (NIHSS) [Brott et al., 1989].

The age effect in daily measures was corrected in the same way as in the database of healthy sleepers. The list of all daily measures and their abbreviations can be found in Table 1.2.

Daily measure	Abbreviation
Fine motor activity test	FMAT2
- number of correctly retraced pixels in a pattern n	FMAT2_cp$n$
- number of successes in pattern n	FMAT2_sp$n$
Reaction time test	RTT
- average reaction time in a trial n	RTT_n
- the minimal reaction time	RTT_min
- the mean reaction time across trials	RTT_mean
Working memory test, forward	WMT_fw
Working memory test, backward	WMT_bw
Lateralised attention network test	LANT
- alerting	LANT_A
- conflict	LANT_C
- orienting	LANT_O
- orienting facilitatory	LANT_OF
- orienting inhibitory	LANT_OI
- right visual field + type (A, C, O, OF, OI)	LANT_RVF_type
- left visual field + type (A, C, O, OF, OI)	LANT_LVF_type
T–MENSTAT questionnaire, before tests	T–MENSTAT_A
- level of energy and motivation	T–MENSTAT_A_1
- level of fatigue and exhaustion	T–MENSTAT_A_2
- level of irritation and frustration	T–MENSTAT_A_3
- level of sleepiness and drowsiness	T–MENSTAT_A_4
T–MENSTAT questionnaire, after tests, part $n = 1, 2, 3, 4$	T–MENSTAT_B_n

Table 1.2: The list of daily measures for the patients after stroke and their abbreviations.

Chapter 2

Sleep modelling and analysis of the sleep process

Sleep is in time continuous process which can be described by a finite set of sleep states during the night. The character of sleep states depends on a chosen sleep model. In this chapter an overview of two sleep models is given. The Rechtschaffen and Kales sleep model (R&K) [Rechtschaffen and Kales, 1968] is widely used in the clinical practice but it produces only discrete representation of the sleep process. On the other hand the Probabilistic sleep model (PSM) [Lewandowski et al., 2012] provides a continuous representation of the sleep and therefore is preferred in this thesis.

In the second part of the chapter the PSM is applied to the database of healthy sleepers as well as to the dataset of patients after an ischemic stroke.

2.1 Rechtschaffen & Kales sleep model

Sleep can be described as a continuous heterogeneous process visiting a finite set of different sleep stages during the night. The R&K sleep model distinguishes six basic sleep stages.

First, it is the *Wake* stage, when a subject doesn't sleep but his or her eyes are closed. The *S1* stage can be characterised as a light sleep. It is similar to the *Wake* stage, but without eye movements under closed eyelids. In the *S2* stage we spend major time during sleep. Presence of so called "spindles" is typical for this stage. Spindles are bursts of an oscillatory brain activity which are visible in the EEG signal [Rechtschaffen and Kales, 1968]. Slow wave sleep is represented by the stages *S3* and *S4*. In contrast to the *S2* stage higher amplitudes and slow waves present in the EEG signal are typical for both of them.

Finally, it is the *REM* stage or the *rapid eye movement*. Muscular and EEG activity similar to wakefulness is typical for this stage. Right before the *REM* occurrence muscular tonic is followed by muscle relaxation. As for its name, quick eye movements behind closed eyelids are typical.

Except of that, R&K model distinguishes also sequences with artefacts (*unscored, UNS*) and segments with movements labelled as *movement time (MT)*.

Nowadays a novel scoring system of the American Academy of Sleep Medicine (AASM) [Iber et al., 2007] is used in the clinical practice. In contrast to R&K the sleep stages *S3* and *S4* are merged into one sleep stage *Slow Wave Sleep* (*SWS*). Stages *S1, S2* and *SWS* (in AASM labelled as *N1, N2, N3*) are also known as *NREM* or *nonREM* stages. Graphical representation of R&K or AASM sleep model is called hypnogram (Figure 2.1).

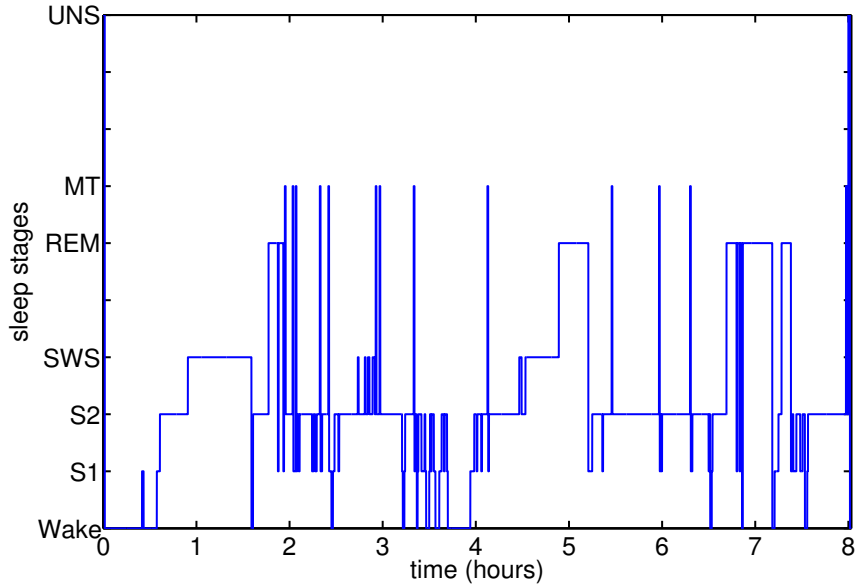


Figure 2.1: Example of a hypnogram of a 22–year–old healthy man. In addition to the standard sleep stages *Wake, S1, S2, SWS* and *REM* also the unscored segments (*UNS*) and segments with movements (*MT*) are depicted.

In both sleep models, the EEG, EMG and other types of biosignals are divided into non–overlapping intervals of the length 30 seconds. Each time window is then classified into one of the above described sleep stages. The staging is done manually by a doctor or by an automatic scoring system, for example Somnolyzer24×7 [Anderer et al., 2005].

In spite of the extensive use of both sleep models in the clinical practice they are under long term criticism, mainly due to the fact that both models produce the sleep process representation with a small number of discrete sleep stages identified at the scale of 30–

second long data segments [Himanen and Hasan, 2000]. In the area of the sleep modelling there is a long term interest for an alternative ways of the sleep representation which successfully overcomes the above mentioned problems. Therefore, in this thesis we focus on the PSM as described in [Lewandowski et al., 2012].

2.2 Probabilistic sleep model

The PSM characterises sleep with probability values of a finite number of sleep states called sleep microstates. Lewandowski et al. [2012] empirically set the number of microstates to 20.

Similarly to R&K or AASM, the PSM is also based on the EEG and EMG signal but the time windows of the length of three seconds are considered. Second difference between the PSM and standard sleep models is in the sleep staging. For each time window probability values for all 20 sleep microstates are estimated in contrast to a discrete assignment into one of the R&K or AASM sleep stages. In other words, each time segment is characterised by a 20–dimensional vector of probability values. Considering the probability values of a given sleep microstate as a function of time we obtain a sleep probabilistic curve.

In the case of the PSM there are no general rules or manuals which can be applied to the PSG recording of a subject. For a given dataset, the PSM have to be trained first and after that it may be applied to a wider set of sleep data.

Before training the model, the data are preprocessed. Lewandowski et al. [2012] considered only the EEG signal from the C3–M2 and C4–M1 electrodes as input for the PSM. First, the EEG signal is divided into non-overlapping three–second long time segments and for each time window, coefficients of an autoregressive model of order M are estimated with the Burg method. Then the time windows are assigned into one of four classes of spindles (*0 = no spindle, 1 = possible spindle, 2 = probable spindle, 3 = certain spindle*) by using a linear discriminant analysis applied to the EEG spectral features, for more details see Lewandowski et al. [2012].

Because of the subject's movement, malfunctioning of electrodes or external factors, the EEG signal includes artefacts. To avoid possible misinterpretation of results the time segments with artefacts are properly detected. Primarily the EEG signal from the C3–M2 electrode is used. When an artefact occurs in a time window of the EEG signal from the C3–M2 electrode, the window is replaced by a corresponding time segment of the EEG signal from the C4–M1 electrode. If it is not possible because of an artefact presence in

the EEG signal from both electrodes, the time window is omitted from further analysis.

Once the data are preprocessed, the model's training begins. Let a denotes the vector of autoregressive coefficients of the length M, $z \in \{1, \ldots, 20\}$ is one of the microstates and $p(z)$ is an unknown probability, that we are in a microstate z at a given time window. A Gaussian mixture model is then estimated in the M–dimensional space of autoregressive coefficients

$$p(a) = \sum_{z=1}^{20} p(z)p(a|z) = \sum_{z=1}^{20} \pi_z \mathcal{N}\left(a|\mu_z, \Sigma_z\right).$$

Here, $\mathcal{N}\left(a|\mu, \Sigma\right)$ denotes the probability distribution function of a normal distribution with mean μ and covariance matrix Σ evaluated at the vector a.

However, the sleep microstates are sometimes missing physiological interpretation. To better understand their structure the PSM estimates their relationships to the standard sleep stages. Let us denote the R&K score for a given three–second long time segment as $r \in \{0 = Wake, 1 = S1, 2 = S2, 3 = SWS, 5 = REM\}$ and $s \in \{0, 1, 2, 3\}$ is a class of spindles. Because a, r and s are independent [Lewandowski et al., 2012], we can write

$$p(a, r, s) = \sum_{z=1}^{20} p(z)p(a|z)p(r|z)p(s|z) = \sum_{z=1}^{20} \pi_z \mathcal{N}\left(a|\mu_z, \Sigma_z\right) p(r|z)p(s|z). \tag{2.1}$$

The conditional probability of a sleep stage $r \in \{0, 1, 2, 3, 5\}$ under the assumption that we are in a microstate $z \in \{1, \ldots, 20\}$ at a given time window is denoted by $p(r|z)$. Similarly, $p(s|z)$ reflects the conditional probability for the spindle occurrence under the assumption that we are in the microstate z. Note that $\sum_{r \in \{0,1,2,3,5\}} p(r|z) = 1$ and $\sum_{s=0}^{3} p(s|z) = 1$ for a given $z \in \{1, \ldots, 20\}$.

The training of the PSM is based on the estimation of unknown probabilities and parameters in (2.1). Because of the latent variable z, the direct use of the maximum likelihood algorithm is not possible. A suitable tool for this case is the Expectation–Maximisation algorithm (*EM* algorithm). For more details about the estimation of the unknown quantities with the *EM* algorithm see Appendix A in [Lewandowski et al., 2012].

Once $p(r|z), r \in \{0, 1, 2, 3, 5\}, z \in \{1, \ldots, 20\}$ are estimated, it is possible to reconstruct probabilistic curves for the sleep stages in the R&K sense and to compare these results with the standard R&K scores.

2.3 Application to the sleep data

2.3.1 The healthy sleepers from the SIESTA database

In the first step we considered the data of healthy sleepers from the SIESTA database. In contrast to the original PSM [Lewandowski et al., 2012], we trained a new version of the model based on the EEG signal from three pairs of electrodes (Fp1–M2, Fp2–M1, C3–M2, C4–M1,O1–M2, O2–M1) and EMG as well. In our previous work [Rošťáková, 2015] we demonstrated that the inclusion of the EMG signal and the EEG signal from more than one pair of electrodes improves the detection of the *REM* related sleep microstates.

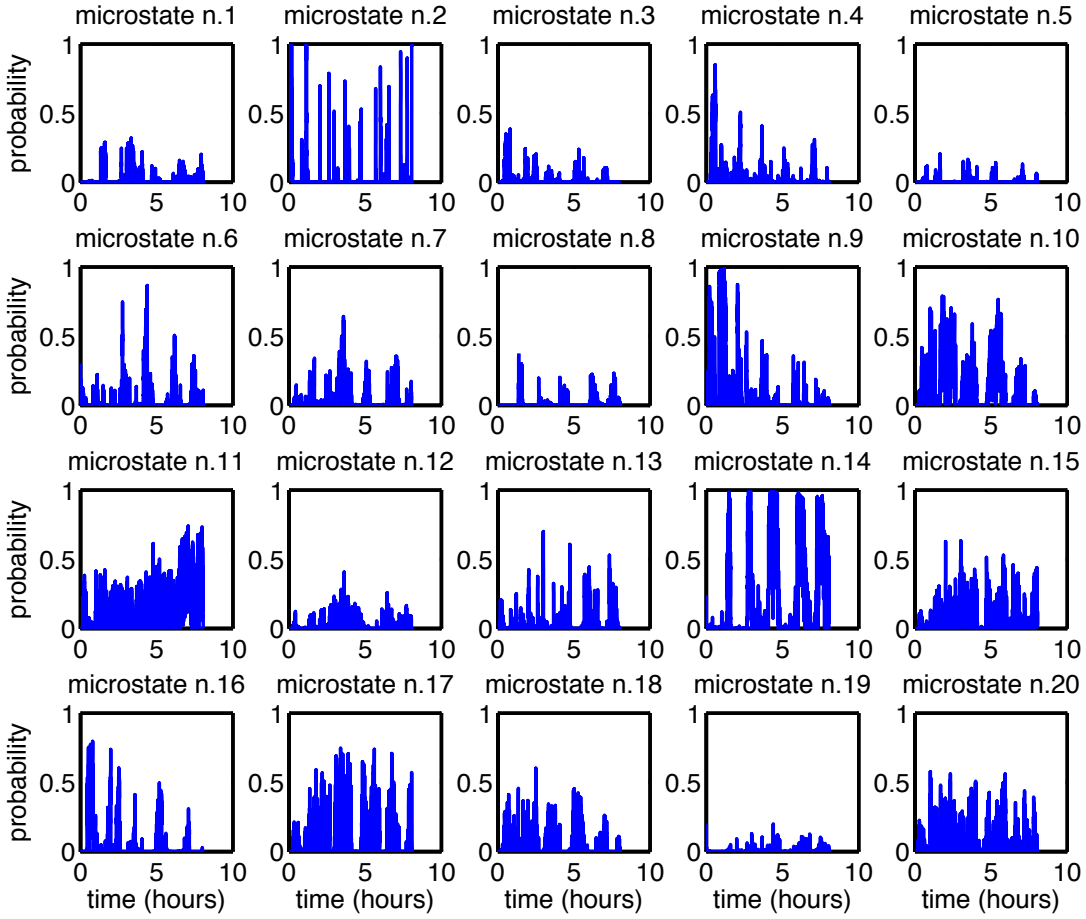


Figure 2.2: An example of the sleep probabilistic curves for 20 microstates of a 42–year–old healthy man.

For the EEG signal, the autoregressive model of order five was chosen, in case of the EMG signal it was the autoregressive model of order two. Autoregressive coefficients of the

EEG signal from three EEG electrodes and the EMG signal were merged into one vector of the length 17 in the following order [Fpx, Cx, Ox, EMG], where x represents either left or artefacts free EEG electrode at the frontal (Fp), central (C) or the occipital (O) spatial site. Details of the choice of an appropriate order of the used autoregressive models for the EEG and EMG signals are given in [Rošťáková, 2015]. Artefacts in the EEG signal for healthy sleepers were detected by the automatic scoring system Somnolyzer24×7 [Anderer et al., 2005]. The remaining part of the training process of the PSM was performed in the same way as in [Lewandowski et al., 2012].

An example of the sleep probabilistic curves for a 42–year–old healthy man is depicted in Figure 2.2 and the sleep probabilistic curves of the standard sleep stages and corresponding R&K scores for the same subject are depicted in Figure 2.3. The similarity weights between the sleep microstates and the standard sleep stages *Wake, S1, S2, SWS* and *REM* are listed in Table 2.1. For example, sleep Microstates 8 and 14 are similar to the *REM* stage with probabilities 74% and 72%, while Microstate 13 lies on a border between the *Wake* stage and the light sleep (*S1* stage).

sleep micro.	*Wake*	*S1*	*S2*	*SWS*	*REM*	sleep micro.	*Wake*	*S1*	*S2*	*SWS*	*REM*
1	0.00	0.02	**0.85**	0.02	0.11	11	0.01	0.05	**0.77**	0.07	0.10
2	**0.73**	0.21	0.04	0.00	0.02	12	0.02	0.12	**0.77**	0.04	0.05
3	0.00	0.00	0.30	**0.70**	0.00	13	**0.45**	**0.41**	0.12	0.00	0.02
4	0.01	0.00	0.25	**0.74**	0.00	14	0.03	0.16	0.09	0.00	**0.72**
5	0.00	0.00	0.35	**0.65**	0.00	15	0.06	0.11	**0.68**	0.00	0.15
6	**0.85**	0.06	0.04	0.00	0.05	16	0.00	0.00	0.04	**0.96**	0.00
7	0.00	0.00	**0.64**	0.36	0.00	17	0.00	0.00	**0.98**	0.02	0.00
8	0.00	0.03	0.23	0.00	**0.74**	18	0.00	0.00	**0.50**	**0.50**	0.00
9	0.03	0.15	**0.75**	0.04	0.03	19	**0.88**	0.07	0.02	0.00	0.03
10	0.00	0.00	**0.88**	0.12	0.00	20	0.01	0.07	**0.64**	0.00	0.28

Table 2.1: The weights representing the similarity between the sleep microstates and the standard sleep stages. The dominant weight for each microstate is shown in bold.

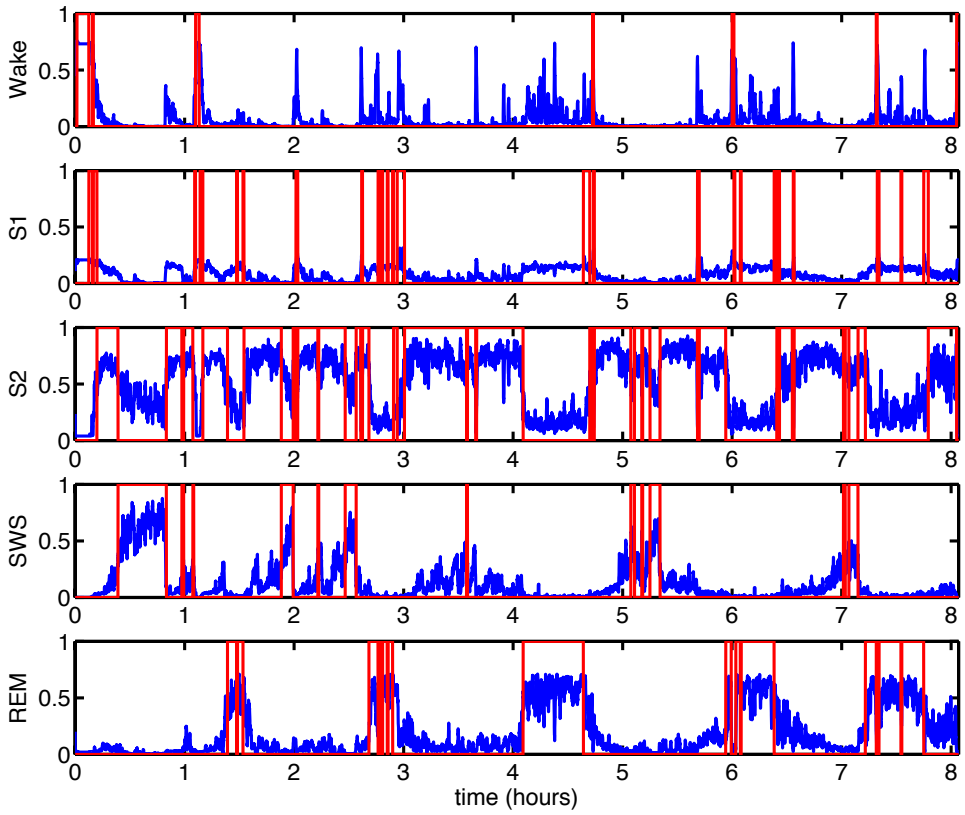


Figure 2.3: An example of the sleep probabilistic curves for sleep stages *Wake, S1, S2, SWS, REM*. The blue curves represent a whole night profile of a 42–year–old healthy man. Corresponding Rechtschaffen and Kales scores [Rechtschaffen and Kales, 1968] are depicted in red.

2.3.2 Patients following an ischemic stroke

The PSM was applied to the set of the PSG recordings of 24 patients following an ischemic stroke for which the results of neurocognitive tests and questionnaire about the sleep quality were available.

In contrast to the subjects from the SIESTA database, the EEG signal from the frontal pair of electrodes was not at disposal. Moreover, because of artefacts presence in the EMG channel in the majority of subjects, the usage of the EMG signal was not possible. Therefore we trained a new version of the PSM only on the EEG signal from the central and occipital pair of electrodes. The order of the autoregressive model representing the EEG signal was set to 10. The presence of artefacts in the EEG signal was automatically detected by the BrainVision Analyser [Brain Products, GmbH, 2013] software and followed by the manual inspection.

In the previous analysis [Rošťáková, 2015] we observed that the PSM trained without the information about the muscle activity is not able to properly distinguish between the *Wake* and *REM* stage at the beginning of night. Therefore in the PSM training process only the relationship between the sleep microstates and the *nonREM* sleep stages *Wake, S1, S2* and *SWS* were estimated (Table 2.2). When reconstructing the sleep probabilistic curves for the standard sleep stages the *REM* stage was artificially added from the R&K scores. In other words, if in a time window the *REM* stage was detected by R&K, the probability for this stage was set to 1 while the probabilities for the *Wake* stage and the *nonREM* stages were set to 0 (Figure 2.4).

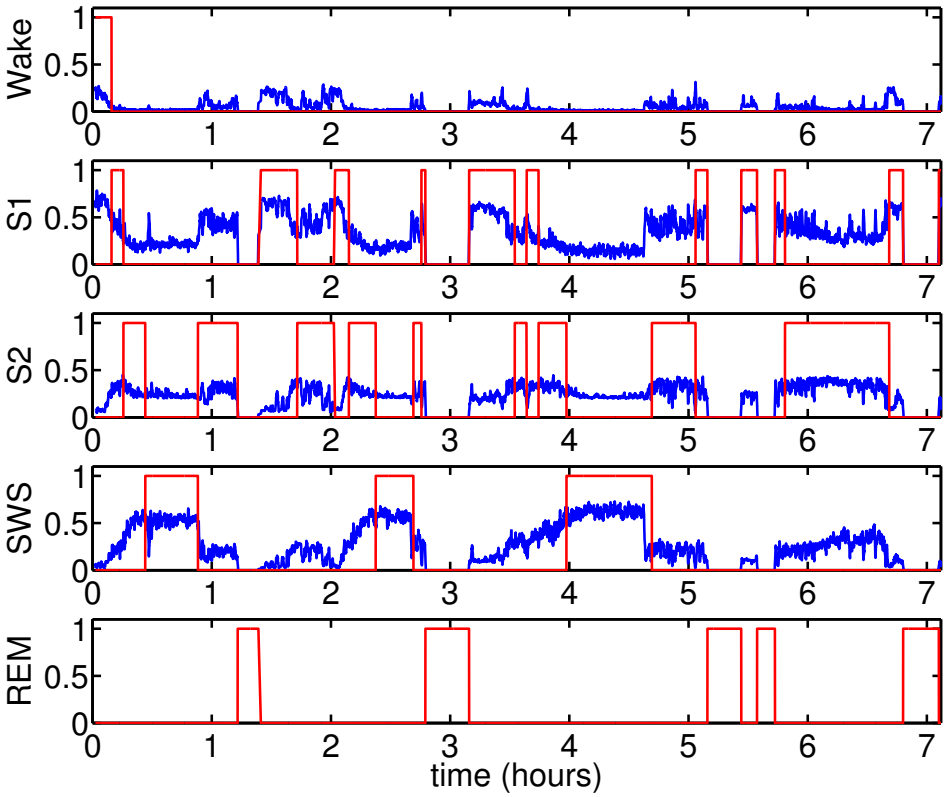


Figure 2.4: An example of the sleep probabilistic curves for the sleep stages *Wake, S1, S2, SWS, REM*. The blue curves represent a whole night profile of a 44–year–old man after an ischemic stroke. Corresponding Rechtschaffen and Kales scores [Rechtschaffen and Kales, 1968] are depicted in red.

Finally, we would like to pay attention to differences in wakefulness and light sleep between patients after stroke and healthy subjects. The healthy sleepers reach higher

probability values for the *Wake* stage, while the probability values for being in the *S1* stage are at most 0.3. In the case of patients after stroke an opposite phenomenon occurred (Figures 2.3 and 2.4). This is mainly caused by similarity between this two stages and different R&K scoring rules used for healthy sleepers and patients after stroke. The R&K scores for the SIESTA database were obtained by the automatic scoring system Somnolyzer [Anderer et al., 2005], while the EEG signal of patients after stroke was scored manually by a trained physician who was likely to put more weight on the *S1 stage*.

sleep microstate	*Wake*	*S1*	*S2*	*SWS*
1	0.37	**0.50**	0.07	0.06
2	0.42	**0.54**	0.02	0.02
3	0.03	0.31	0.22	**0.44**
4	0.02	0.28	**0.42**	0.28
5	0.22	**0.68**	0.08	0.02
6	0.01	**0.44**	**0.44**	0.11
7	0.25	**0.65**	0.07	0.03
8	0.00	0.05	0.20	**0.75**
9	0.00	0.24	0.19	**0.57**
10	0.05	**0.59**	0.25	0.11

sleep microstate	*Wake*	*S1*	*S2*	*SWS*
11	0.01	0.15	0.19	**0.65**
12	0.04	0.21	**0.44**	0.31
13	0.01	**0.44**	0.25	0.30
14	0.13	**0.81**	0.04	0.02
15	0.11	**0.76**	0.12	0.01
16	0.04	**0.51**	0.34	0.11
17	0.00	0.21	**0.49**	0.30
18	0.00	0.08	0.24	**0.68**
19	0.00	0.02	0.12	**0.86**
20	0.12	**0.71**	0.08	0.09

Table 2.2: The weights representing the similarity between the sleep microstates and the standard sleep stages. The dominant weight for each microstate is shown in bold.

Chapter 3

Functional data clustering

The sleep probabilistic curves can be considered as functional data. Levitin et al. [2007] defines functional datum as a set of measurements along a continuum, for example time. From some point of view they are similar to the longitudinal data. The difference is that the functional data are observed in pairs (t_i, x_i) where t_i is the i^{th} time point in which the functional datum was observed and x_i is observed value.

In the functional data analysis (FDA) dividing a given set of curves into subgroups with common features is helpful for better understanding of an existing structure and variability in data. We speak about functional cluster analysis.

The structure of the chapter is as follows. First, the definition of functional space, corresponding inner product and the measures of distance between curves are given. Then chosen methods for cluster analysis and rules for finding an appropriate number of clusters are described.

3.1 Functional space

Let consider a functional space $\mathcal{F}$ of all square–integrable, real valued functions defined on a time interval T. For simplicity we assume $T = [0, 1]$, but the majority of methods described below are able to work with a general time interval. This functional space is the well–known $L^2[0, 1]$ space. In the text below both terms "curve" and "function" will be used to label the elements of $L^2[0, 1]$.

The inner product of two functions $X, Y \in L^2[0, 1]$ is defined by the formula

$$\langle X, Y \rangle = \int_0^1 X(t)Y(t)dt. \tag{3.1}$$

Consequently the L^2 norm of a function X is defined as $\|X\| = \sqrt{\langle X, X\rangle}$.

However, $L^2[0,1]$ is an infinite dimensional functional space which may cause problems in a statistical analysis. Therefore it is appropriate to map its elements into a finite dimensional vector space using a dimensionality reduction method. The main idea is to express each element of the space $\mathcal{F}$ by a linear combination of a finite set of basis functions $\varphi_1, \ldots, \varphi_K$

$$X(t) = \sum_{i=1}^{K} a_i \varphi_i(t)$$

and then to work with the vector of coefficients $a = (a_1, \ldots, a_K)^T$.

3.2 Distance measures between curves

Let consider a pair of curves X, Y defined on a closed time interval $T = [0,1]$. In this section we focus on three distance measures which will play important role in the following chapters. For more examples see [Montero and Vilar, 2014].

In a vector space the Minkowski distance between two vectors is very popular. Its analogue for curves is

$$d_n(X,Y) = \left(\int_T \Big(X(t) - Y(t) \Big)^n dt \right)^{\frac{1}{n}}. \tag{3.2}$$

Two curves are similar according to this measure if the area between them is small. With $n = 2$ the formula (3.2) is called the L^2 distance between curves X and Y.

A modified version of the criterion used in [Sangalli et al., 2010] forms another type of distance criterion

$$d_\rho(X,Y) = 1 - \frac{\int_T X'(t) Y'(t) dt}{\sqrt{\int_T \left(X'(t)\right)^2 dt} \sqrt{\int_T \left(Y'(t)\right)^2 dt}}. \tag{3.3}$$

Here, $d_\rho(X,Y) \in [0,2]$ and $d_\rho(X,Y) = 0$ if and only if there exists a finite sequence $s_0 = 0 < s_1 < \cdots < s_m = 1$ which divides the time interval T into m non–overlapping subintervals and

$$\begin{aligned} X(t) = a_i Y(t) + b_i, \quad & t \in [s_{i-1}, s_i), \\ & a_i \in \mathbb{R}^+, b_i \in \mathbb{R}, \quad i = 1, \ldots, m. \end{aligned}$$

The third criterion

$$d(X,Y) = \sqrt{\int_T \Big(\operatorname{sign}(X'(t)) \sqrt{|X'(t)|} - \operatorname{sign}(Y'(t)) \sqrt{|Y'(t)|} \Big)^2 dt}$$

is a special case of the Fisher–Rao metric. Here, sign represents the function *signum*

$$\operatorname{sign}(x) = \begin{cases} -1, & x < 0, \\ 0, & x = 0, \\ 1 & x > 0. \end{cases}$$

Advantages of this distance will be described in Section 4.3.3.

3.3 Methods for curve clustering

Let $\{X_1, \ldots, X_N\}$ be a set of curves defined over the time interval T which should be divided into K clusters. Jacques and Preda [2014] distinguish three main approaches for the functional data clustering.

The first one is called the *raw data clustering*. In real data representation, only a discrete observation for each curve $X_i,\ i = 1, \ldots, N$ in a finite set of time points (knots) $\{t_{i1}, \ldots, t_{im_i}\} \in T$ is available. When the curves are observed over the same set of time points, many methods developed for clustering vectors like the k–means or k–medoids are appropriate also in the functional case.

The second approach is based on the idea, that the curves $X_1, \ldots, X_N$ belonging to a possibly infinite dimensional functional space $\mathcal{F}$ may be expressed as a linear combination of a functional basis $\varphi_1, \ldots, \varphi_L$

$$X_i(t) = \sum_{j=1}^{L} a_{ij}\varphi_j(t),\ \ t \in T$$

The choice of the basis functions depends on the character of the data, for example Legendre polynomials form functional basis for functions observed over interval $[-1, 1]$ or for the square–integrable functions of the $L^2[0, 1]$ space the Fourier basis is appropriate. Another example are splines [Wahba, 1990]. The functional basis can be reconstructed also directly from the data by the Functional principal component analysis (FPCA) [Yao et al., 2003], for more details see Section 7.1.

Now, each curve $X_i,\ i = 1, \ldots, N$ is represented by its vector of coefficients $a_i = (a_{i1}, \ldots, a_{iL})^T,\ i = 1, \ldots, N$ so the standard clustering tools for finite dimensional data can be applied.

The third way of clustering functional data consists of standard methods like the k–means method or hierarchical clustering in which the vector distance is replaced by appropriate distance measure for curves.

In the analysis of sleep data we prefer to use the modified k–means or k–medoids methods following the idea of the later two approaches.

3.4 Number of clusters

The important part of clustering methods, for example the k–means or k–medoids algorithm, is the selection of the number of clusters. However, because of a strong individuality of sleep, which can be present in each subject's curve, to choose a proper number of clusters is a challenging task. Kodinariya and Makwana [2013] present an overview of criteria for choosing an appropriate number of clusters K in multivariate data. Selected criteria can be applied to functional data. Following these lines, in this thesis we used the

- **rule of thumb**, $K \approx \sqrt{\frac{N}{2}}$, where N is the cardinality of the dataset.
- **elbow diagram**

  Let consider a cost function with the number of clusters k as its argument representing the homogeneity of the constructed clusters

$$L(k) = \frac{1}{N} \sum_{i=1}^{k} \sum_{j:X_j \in C_i} \int_T (X_j(t) - \mu_i(t))^2 \, dt, \tag{3.4}$$

$$\mu_i(t) = \frac{1}{|C_i|} \sum_{j:X_j \in C_i} X_j(t),$$

  where C_i represents the i^{th} cluster $i = 1, \ldots, k$, μ_i is its centroid and $|C_i|$ represents its cardinality. We call this cost function the L–criterion. Then values of the L–criterion for the chosen clustering method are plotted as a function of the number of clusters. This dependance is visually inspected and searched for a rapid drop followed by a visible plateau. This rapid drop point is then selected for setting the number of cluster K.

- **average silhouette**

  The *silhouette* [Rousseeuw, 1987] represents tightness and separation of each cluster. It reaches values from the interval $[-1, 1]$ and shows how well a curve lies within the cluster and which curves lay in between clusters. Let consider the curve X_j which belongs to the i^{th} cluster C_i, $i = 1, \ldots, k$. Then its silhouette $sil(k, X_j)$ is defined as a normalised difference between the within–cluster homogeneity and separation

from the rest of the dataset

$$sil(k, X_j) = \frac{H_B^k(j) - H_W^k(j)}{\max\{H_B^k(j), H_W^k(j)\}},$$

where $H_W^k(j)$ is an average L^2 distance between X_j and other curves in the i^{th} cluster. The second term $H_B^k(j)$ is an average distance between X_j and all curves from the nearest cluster in the L^2 distance sense. When $sil(k, X_j)$ reaches its maximum, $sil(k, X_j) = 1$, it means that X_j is well–clustered. Minimum value of $sil(k, X_j)$ implies that X_j is classified into incorrect cluster and $sil(k, X_j) = 0$ means that X_j lies on a border between two clusters. We can also define the *average silhouette* (AS)

$$\overline{sil}(k) = \frac{1}{N} \sum_{i=1}^{N} sil(k, X_j)$$

which varies with k. Naturally, $\overline{sil}(k)$ is approximately 1 when k is close to N. Therefore we restrict the value of k to be between 2 and $\frac{N}{2}$. The optimal number of clusters K is then chosen as the argument of the local maximum of $\overline{sil}$. If there is no visible global maximum of $\overline{sil}$, then the optimal number of clusters is set to that k after which $\overline{sil}$ reaches approximately constant values.

In general, there doesn't exist a single criterion which would be ideal for all data. This is because each rule classifies quality of clustering in a different way. In this thesis we computed the optimal number of clusters according to different criteria and we chose the value k which occurred the most frequently.

3.5 Relationship between the structure of clusters and daily measures

The focused objective of our current research is to identify specific sleep profiles (sleep biomarkers) associated with selected physiological aspects of sleep. An important property of such sleep biomarkers would be their relationship with different physiological, demographic or daily life measures. This may include physiological factors as blood pressure, pulse rate or others, results of questionnaires about subjective sleep quality, mood, drowsiness or results of neurophysiological tests focused on attention, fine motor activity or short–term memory [Rosipal et al., 2013].

A standard approach how to achieve this goal is to extract one–dimensional sleep variables from the R&K or PSM sleep models and then to compute their correlation with

a daily measure represented by one of the tests or questionnaires listed in Table 1.1 [Rosipal et al., 2013; Škoviera et al., 2014].

The advantage of the PSM is that the sleep probabilistic curves describe the whole dynamic of the sleep process. However, the one–dimensional sleep characteristics extracted from the PSM may miss some information about the sleep dynamics. Therefore we prefer to use the whole sleep probabilistic curve information when detecting relationships between the sleep structure and daily life behaviour.

One way of the sleep structure analysis is the transformation of possibly infinite dimensional sleep probabilistic curves into a finite dimensional vector space by using the FPCA (Section 7.1). Now, each curve is represented by a vector and standard tools from multivariate statistics can be applied. This approach was studied in [Rošťáková and Rosipal, 2018].

Another approach is to divide the set of all sleep probabilistic curves into subgroups according to the similarity in their shape. We speak about cluster analysis of curves. An overview of existing methods and discussion of problems which may occur during curves clustering are described in the next chapters. Once the clusters are formed (by an arbitrary method) we aim to detect, whether differences in the sleep profiles between clusters are mirrored also in differences in daily measures. In general, the one–way ANOVA test can be used for this purpose. However, because of presence of outlier values in the set of daily measures we prefer to use its non–parametric version represented by the Kruskal–Wallis test.

Chapter 4

Time synchronisation of curves

In the previous chapter we described several approaches for the curves clustering. Once the sleep probabilistic curves are assigned into clusters, it is possible to test whether there is a significant difference in daily measures between clusters. Our preliminary study [Rošťáková, 2015] showed existing relationships between the sleep structure and age, subjectively scored sleep quality or the level of drive and drowsiness in the morning. However, we hypothesised, that some relationships remained hidden because of the curves misalignment.

We say, that two curves are similar, but misaligned in time, if they have a common overall shape, but their important features like local maxima, minima or zero crossings occur at different time points. Therefore, the criteria from Section 3.2 may reach high values and consequently the clustering techniques will consider these curves as dissimilar and assign them into different clusters.

To avoid this problem Ramsay and Silverman [2005] recommend to synchronise curves before further analysis. In the literature the curves synchronisation problem is also known as the curves alignment or curves registration and several methods were developed to solve this problem.

At the beginning of this chapter problems of cluster analysis applied to in time misaligned curves is demonstrated on artificially generated data. Then the curves alignment is formulated in a rigorous mathematical way. Finally, a set of chosen approaches for the curves registration is described in details and applied to the database of healthy sleepers.

4.1 Problem of clustering of in time misaligned curves

Let consider five curves $\mu_1, \mu_2, \mu_3, \mu_4, \mu_5$ defined on the $[0, 1]$ time interval

$$\begin{aligned}
\mu_1(t) &= 1.2e^{-20(t-0.7)^2} - e^{-50(t-0.45)^2} + 1.2e^{-100(t-0.3)^2} - 1.2e^{-150(t-0.2)^2} + \\
&\quad + e^{-200(t-0.15)^2} - 1.76e^{-300(t-0.05)^2}, \\
\mu_2(t) &= \mu_1(t-0.1) - 0.6\mathcal{N}\left(\frac{t-0.8}{0.01}\middle|0,1\right) + \mathcal{N}\left(\frac{t-0.2}{0.2}\middle|0,1\right), \\
\mu_3(t) &= \mu_1(t+0.1) + \mathcal{N}\left(\frac{t-0.8}{0.01}\middle|0,1\right) - 0.4\mathcal{N}\left(\frac{t-0.2}{0.02}\middle|0,1\right), \\
\mu_4(t) &= \frac{1}{0.25}\mathcal{N}\left(\frac{2t-1}{0.25}\middle|0,1\right), \\
\mu_5(t) &= \frac{1}{0.1}\mathcal{N}\left(\frac{2t-0.3}{0.1}\middle|0,1\right) + \frac{1}{0.1}\mathcal{N}\left(\frac{2t-1.5}{0.1}\middle|0,1\right) \qquad t \in [0,1]. \qquad (4.1)
\end{aligned}$$

Here, $\mathcal{N}(t|0,1)$ represents the probability distribution function of the normalised normal distribution at the point t. These curves will serve as template curves. They were chosen to be similar on short subintervals, but with different profiles when considering the whole time interval $[0, 1]$.

For the first three template functions we consider a sample of 10 curves, for μ_4 and μ_5 this number was set to 20 curves. Curves $X_i,\ i = 1, \ldots, 70 = 3 \times 10 + 2 \times 20$ were generated by the following formula

$$X_i(t) = a_i \mu(g_i(t))$$

where μ represents one of the template curves, the multiples $a_i, i = 1, \ldots, N$ were obtained as random samples from the normal distribution $\mathcal{N}(1, 0.1)$. The time transformations g_i were simulated as

$$g_i(t) = c(t + \alpha_{2i} e^{t-\alpha_{1i}} + b)$$

The normalising constants b and c guarantees that $g_i(0) = 0$ and $g_i(1) = 1$. Parameters α_{1i} and α_{2i} were generated from the normal distribution $\mathcal{N}(0, 4)$ and $\mathcal{N}(1, 10^{-4})$ respectively. An example of the data and template curves is depicted in Figure 4.1.

After applying the k–means algorithm with the L^2 distance several curves were assigned into incorrect cluster (Figure 4.2). The time misalignment caused that the curves originally generated from the template curves μ_3 and μ_4 were assigned into the same cluster although

their profile is different. On the other hand, the curves generated from the template μ_5 were divided into two clusters.

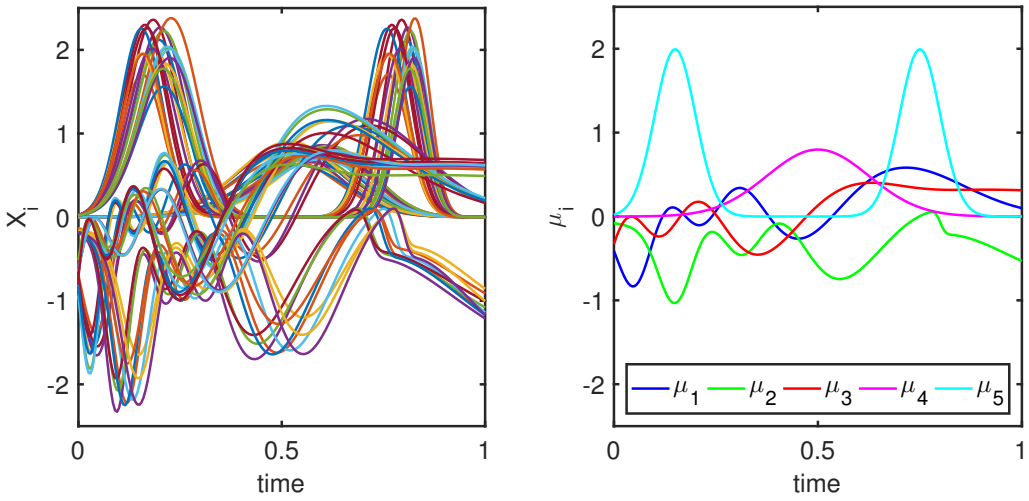


Figure 4.1: Example of simulated data (left) generated from five template curves defined in (4.1) (right).

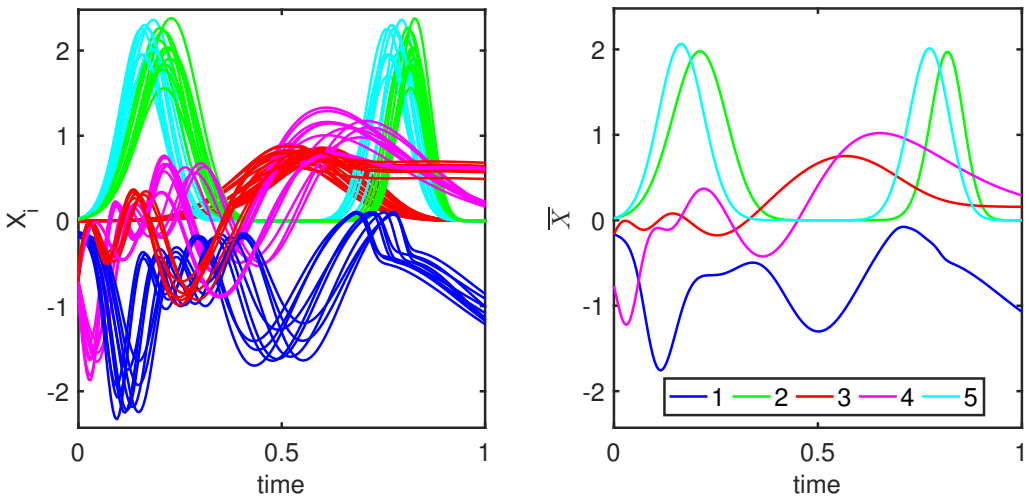


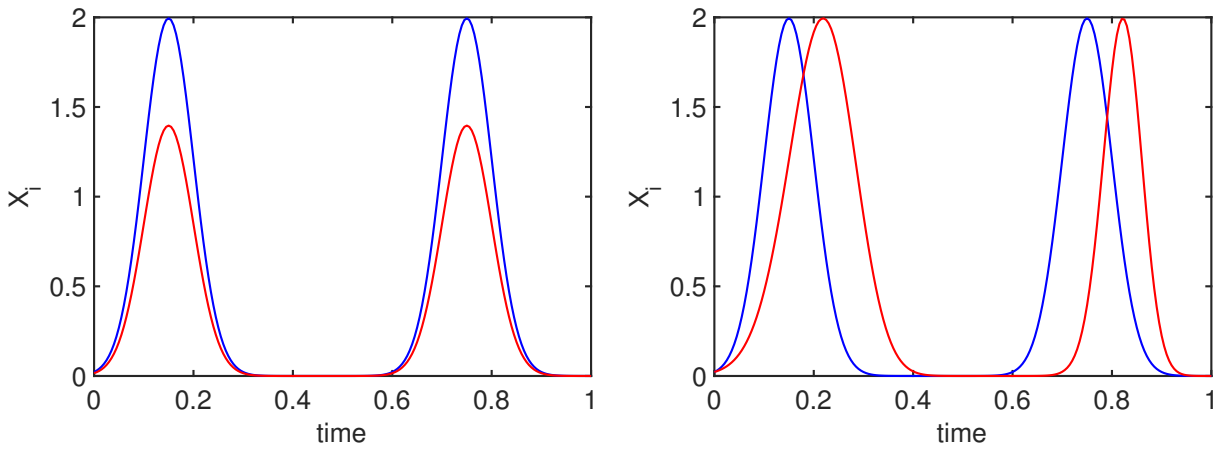
Figure 4.2: Cluster analysis of the artificial data from Figure 4.1. Five clusters were generated by the k–means algorithm (left), cluster representatives were computed as an average curve of each cluster (right).

Except of clustering into incorrect group, also another problem occurred. Once the clusters are formed, we would like to characterise them by one curve with the cluster–typical

shape. However, when the curves are not in time synchronised the cluster representatives can miss important features. For example, the peak of curves from the cluster 1 located in the time interval $[0.6, 0.8]$ (Figure 4.2, left) is missing in the cluster representative (Figure 4.2, right). All these problems stress the importance of applying the curve synchronisation step before running cluster analysis.

4.2 Definition of the curve alignment problem

Curves with similar profile differ in a few ways. Ramsay and Silverman [2005] distinguish two basic types of variation which can be present in the functional data – amplitude and time variability. Two curves vary in the amplitude, if their important features like local maxima or minima occur at the same time, but they have different height (Figure 4.3a). Time variation means that common features occur at different time points without considering their height (Figure 4.3b). In this analysis we focus on the second type of variability and methods for its removing or suppression.



(a) amplitude variability

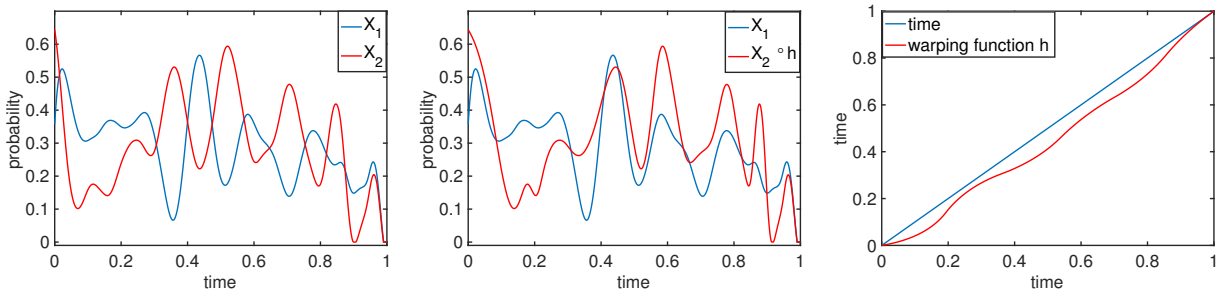
(b) time variability

Figure 4.3: Example of two curves with similar profile varying in amplitude (left) and time (right).

Let consider a pair of curves X_1, X_2 which are defined over a closed time interval T. Without loss of generality we assume $T = [0, 1]$. To register or temporally align a pair of curves X_1, X_2 means to find a function $h : T \to T$ from the set $\mathcal{H}$ of all strictly increasing transformations of the time interval T such that

$$h \in \text{argmin}_{h^\star \in \mathcal{H}} S\left(X_1, X_2 \circ h^\star\right), \tag{4.2}$$

where S is a measure of distance between two curves X_1, X_2; for example one of the measures defined in Section 3.2. The time transformation $h^\star$ is called the warping function. An example of two in time misaligned smoothed sleep probabilistic curves, their aligned version by using the criterion (3.2) and the corresponding warping function is depicted in Figure 4.4.



(a) original curves (b) aligned curves (c) warping function

Figure 4.4: An example of two smoothed sleep probabilistic curves varying in time (left), their aligned versions by using the criterion (3.2) (middle) and the corresponding warping function (right). The original time was transformed into the interval [0,1].

There are many techniques developed for solving the problem (4.2). As we will see in the following sections, they differ in the type of criterion S and in the algorithm used for finding the appropriate warping function.

In addition to the assumption of strictly increasing warping function h we consider three additional restrictions which appear frequently in literature:

- *h is differentiable* – the assumption is necessary when a cost function includes also derivative of the warping function h

- *the assumption of the common start and end point* – the warping function h should begin and end at the same time point as the real time. In other words

$$h(0) = 0, \qquad h(1) = 1.$$

- *restriction to the distance between the real time and the warping function* – the goal of a registration method is to transform curves in time such that they are as similar as possible. However, allowing the warping function to run too far from the real time may result in rapid distortions of the curves profiles and consequently we can obtain close to ideal alignment of possibly dissimilar curves. Therefore it is appropriate to

add a constraint to the distance between the real time and warping function h into the cost function, for example in the form

$$\lambda \int_T \left(h^{'}(t) - 1\right)^2 dt.$$

When the penalty parameter λ is small the warping function h can be far from the physical time which can cause that distant segments of curves are aligned to each other. On the other hand, large values of the penalty parameter restrict h to be close to the real time which can lead to insufficient alignment. Appropriate λ can be found for example by cross–validation. More examples of restrictions to the distance between the warping function h and the real time can be found in [Srivastava and Klassen, 2016, Chapter 8].

4.3 Current state of the curves alignment problem

In the literature there are many methods developed for solving the curves misalignment problem. We can divide them into several groups according to the form of the warping functions and the character of alignment.

The first group includes methods which consider only linear time transformations in the form $h(t) = at + b, \;\; a > 0, b \in \mathbb{R}$, for example the Constant time shift method [Ramsay and Silverman, 2005], the k–mean alignment for the curve clustering [Sangalli et al., 2010] or the Joint probabilistic curve clustering and alignment [Gaffney and Smyth, 2005]. Of course, in this case the condition of the common start and end point is not taken into account.

In practice, the curves misalignment has usually a nonlinear character, therefore the majority of alignment methods are able to work also with nonlinear time transformations under the condition of the common start and end point. This set of alignment methods can be further divided into methods which align each curve to a common target curve and methods based on alignment of each pair of curves separately. The first type of methods is represented by the Self–modelling time warping method [Gervini and Gasser, 2004], Curve registration using fractional polynomials [Bugli et al., 2005], the Registration to the principal components [Kneip and Ramsay, 2008], the Elastic time warping method [Tucker et al., 2013] or the Curves alignment by moments [James, 2007]. The second part of methods includes the Pairwise curve synchronisation [Müller and Tang, 2008] or the Dynamic time warping [Wang and Gasser, 1997].

Following our practical experience with the time alignment of the sleep probabilistic curves in this thesis we focus on

- the Self–modelling time warping (SMTW) [Gervini and Gasser, 2004],
- the Elastic time warping (ETW) [Tucker et al., 2013],
- the Pairwise curve synchronisation (PCS) [Müller and Tang, 2008].

4.3.1 Self–modelling time warping

Let consider a set of curves $X_i : T \to \mathbb{R},\ i = 1, \ldots, N$ defined on the time interval $T = [0, 1]$. Gervini and Gasser [2004] assume that the set follows the model

$$X_i(h_i(t)) = a_i\mu(t) + \varepsilon_i(t), \quad t \in T.$$

Here, μ is a differentiable target function. Random errors $\varepsilon_1, \ldots, \varepsilon_N$ are modelled as independent identically distributed (*iid*) stochastic processes with a zero mean function. Coefficients $a_1, \ldots, a_N$ represent random multiples of the target function μ. They are assumed to be *iid* with $\mathrm{E}(a_i) = 1,\ i = 1, \ldots, N$. As in the previous sections $h_i : T \to T,\ i = 1, \ldots, N$ represent strictly increasing warping functions satisfying the condition of the common start and end point. Moreover we assume, that they are differentiable.

The main goal of the SMTW method is to find warping functions $h_1, \ldots, h_N$, the target function μ and the coefficients $a_1, \ldots, a_N$ which minimise the criterion

$$S_{SMTW}\left(\mu, a_1, \ldots, a_N, h_1, \ldots, h_N\right) = \sum_{i=1}^{N} \int_T \left(X_i(h_i(t)) - a_i\mu(t)\right)^2 h_i'(t)dt. \tag{4.3}$$

If the warping functions $h_i, i \in \{1, \ldots, N\}$ are known, Gervini and Gasser [2004] derived explicit formulas for $\widehat{\mu}$ and $\widehat{a}_i, i \in \{1, \ldots, N\}$. This leads to the iterative algorithm for minimisation of (4.3),

- choose initial estimates $\widehat{h_1}^0, \ldots, \widehat{h_N}^0$,
- in the $k^{th}, k \geq 1$ step of iteration
  - find $\widehat{\mu}^k, \widehat{a_1}^k, \ldots, \widehat{a_N}^k$ by using $\widehat{h_1}^{k-1}, \ldots, \widehat{h_N}^{k-1}$, obtained in the previous iteration and formulas derived in [Gervini and Gasser, 2004],
  - update estimates $\widehat{h_1}^k, \ldots, \widehat{h_N}^k$ of warping functions by the Newton–Rhapson algorithm.

Repeat these steps until the maximal number of iterations is not reached or until (4.3) is under a chosen threshold.

Gervini and Gasser [2004] expressed the warping function as a sum of the real time and a linear combination of B–splines $B_1, \ldots, B_L$ [Wahba, 1990] of a chosen order

$$h_i(t) = t + \sum_{j=1}^{K} s_{ij} \sum_{l=1}^{L} p_{jl} B_l(t), \quad i = 1, \ldots, N, \;\; t \in T.$$

The coefficients $p_{jl}, j = 1, \ldots, K; l = 1, \ldots, L$ and $s_{ij}, i = 1, \ldots, N; j = 1, \ldots, K$ are constrained such that the condition of the common start and end point is satisfied. However, the constraints do not guarantee that the warping functions $h_1, \ldots, h_N$ are strictly increasing, only nondecreasing. The authors solve this problem by iteratively repeating the Newton–Rhapson step in the algorithm until the estimated warping functions are strictly increasing (see the MATLAB code with the implemented SMTW method which is available at the author's webpage [Gervini, 2004]). In the original MATLAB code 20 iterations are set as a default value for checking monotonicity. Gervini and Gasser [2004] claim that in the majority of datasets this number of iterations is sufficient to obtain a strictly increasing warping function. However, as we will show in Section 4.4 there are data where even a larger number of repetitions doesn't lead to strictly increasing warping functions.

In the light of these findings we need to consider a restriction to the warping function h. A nondecreasing function main contain segments with a constant value over some short intervals and therefore its first derivative is zero there. Because of that we modified the cost function (4.3) by adding the following penalty term

$$\begin{aligned} S^{\star}_{SMTW}(\mu, a_1, \ldots, a_N, h_1, \ldots, h_N) = S_{SMTW}(\mu, a_1, \ldots, a_N, h_1, \ldots, h_N) \\ + \lambda \sum_{i=1}^{N} \int_T \left(\frac{1}{h_i'(t)} - 1 \right)^2 dt. \end{aligned} \tag{4.4}$$

The penalty avoids the estimate of the warping functions with the first derivative being close to 0 because the expression $\frac{1}{h'(t)}, t \in T_0$ is large when $h(t), t \in T_0$ is approximately constant over the interval $T_0 \subset T$. Moreover, to avoid alignment of very distant segments of curves, the term $\frac{1}{h'(t)}$ should not go far from 1. The constant λ influences the weight of the penalty. The main idea of the algorithm doesn't change by adding the penalty term.

With the aim to find optimal penalty weight, we varied λ between 0 and 0.5 ($\lambda > 0.5$ led to poor or no alignment) and for each case the mean squared error MSE of aligned

curves

$$MSE(X_1 \circ h_1, \ldots, X_N \circ h_N) = \frac{1}{N} \sum_{i=1}^{N} \int_T \left((X_i \circ h_i)(t) - \overline{X}(t) \right)^2 dt,$$

$$\overline{X}(t) = \frac{1}{N} \sum_{i=1}^{N} (X_i \circ h_i)(t)$$

was computed. As expected, the lowest MSE was obtained for $\lambda = 0$. The optimal λ was then selected as the value after which the MSE was approximately constant or only slightly increasing.

4.3.2 Pairwise curve synchronisation

The PCS method [Müller and Tang, 2008] solves the curves misalignment problem from a different point of view. As for its name, this method aligns all possible pairs of curves separately and after that for each curve a global warping function is estimated.

Let consider a set of curves $X_1, \ldots, X_N$ defined on the time interval $T = [0, 1]$. For each pair of curves $X_i, X_j, i \neq j$ our main objective is to identify the warping function h_{ij} which minimises the following cost function with the restriction term controlling the distance between the real time and the warping function

$$S_{PCS}(X_i, X_j, h_{ij}) = \int_T \left(X_i(h_{ij}(t)) - X_j(t) \right)^2 dt + \lambda \int_T \left(h_{ij}(t) - t \right)^2 dt. \tag{4.5}$$

Here we would like to highlight that $h_{ij} \neq h_{ji}^{-1}$ in general, while S_{PCS} is not symmetric to a random warping. Exact methodology for finding appropriate value of the penalty parameter λ can be found in [Müller and Tang, 2008; Müller, 2012].

Each warping function $h_{ij},\ i, j = 1, \ldots, N$ is modelled in the following way

$$\begin{aligned}
h_{ij}(t) &= \sum_{k=0}^{p+1} \tau_{ijk} A_k(t), \\
A_k(t) &= \frac{t - a_{k-1} \mathrm{I}_{[a_{k-1}, a_k)}(t)}{a_k - a_{k-1}} - \frac{t - a_{k+1} \mathrm{I}_{[a_k, a_{k+1})}(t)}{a_{k+1} - a_k}, \quad k = 0, \ldots, p, \\
A_{p+1} &= \frac{t - a_p \mathrm{I}_{[a_p, a_{p+1})}(t)}{a_{p+1} - a_p},
\end{aligned}$$

where $\tau_{ij0}, \ldots, \tau_{ij(p+1)}$ are unknown coefficient and $0 = a_0 < a_1 < \cdots < a_p < a_{p+1} = 1$ is a set of knots which divide the interval $[0, 1]$ into $p + 1$ subintervals of an equal length. The indicator functions $\mathrm{I}_{[a_k, a_{k+1})}, k = 0, \ldots, p$ are defined by the following formula

$$\mathrm{I}_{[a_k, a_{k+1})}(t) = \begin{cases} 1, & \text{if} \quad t \in [a_k, a_{k+1}), \\ 0, & \text{if} \quad t \notin [a_k, a_{k+1}). \end{cases}$$

To satisfy the condition of the common start and end point $\tau_0 = 0$ and $\tau_{p+1} = 1$. Furthermore the assumption $\tau_{j+1} > \tau_j,\ \ j = 0, \ldots, p$, guarantees that the estimated warping functions are strictly increasing.

Once $h_{ij},\ \ i, j = 1, \ldots, N$ are estimated, Müller and Tang [2008] define the global warping function $h_i(t)$ for each curve $X_i, i = 1, \ldots, N$

$$h_i(t) = \left(\frac{1}{N} \sum_{j=1}^{N} h_{ji}(t) \right)^{-1}, \tag{4.6}$$

where the upper-script -1 represents the inverse function.

Truncated averaging procedure forms a more robust alternative to (4.6). It means that only a pair of curves whose L^2 distance after the warping process is less than a given constant are used in the process of estimation of the global warping function

$$h_i(t) = \left(\sum_{i=1}^{N} \frac{h_{ji}(t) \mathrm{I}_{d_2^2(X_j \circ h_{ji}, X_i) < \epsilon}}{\sum_{l=1}^{N} \mathrm{I}_{d_2^2(X_j \circ h_{jl}, X_l) < \epsilon}} \right)^{-1},$$

where d_2^2 is the distance measure defined in (3.2). More details about the truncated version of PCS can be found in [Müller and Tang, 2008; Tang and Müller, 2009]. A MATLAB script which implements both, the original and truncated PCS method, is at disposal in the MATLAB PACE toolbox [Müller, 2012].

4.3.3 Elastic time warping

The ETW method [Tucker et al., 2013] belongs to a set of registration methods which align a set of curves to a target curve. Let consider a set $X_1, \ldots, X_N$ of absolutely continuous functions defined over the time interval $T = [0, 1]$. In addition to the condition of the common start and end point, Tucker et al. [2013] assume that the warping function $h \in \mathcal{H}$ is a diffeomorphism which means that h is a differentiable function and there exists its differentiable inversion function.

For each $X_i,\ \ i = 1 \ldots, N$ Tucker et al. [2013] define the *square root slope function* (SRSF) by the formula

$$q_{X_i}(t) = \mathrm{sign}(X_i'(t)) \sqrt{|X_i'(t)|}, \quad t \in T.$$

A curve X_i can be simply reconstructed from its SRSF q_{X_i} by

$$X_i(t) = X_i(0) + \int_0^t q_{X_i}(t) |q_{X_i}(t)| dt. \tag{4.7}$$

The difference between ETW and other registration methods is in the definition of a distance between two curves X_i, X_j

$$d_{ETW}(X_i, X_j) = \sqrt{\int_T \left(q_{X_i}(t) - q_{X_j}(t)\right)^2 dt}, \tag{4.8}$$

which is the standard L^2 norm of the corresponding SRSFs. The advantage of (4.8) is its invariance to a random warping $h^\star \in \mathcal{H}$

$$\begin{aligned} d^2_{ETW}(X_i \circ h^\star, X_j \circ h^\star) &= \int_T \left(q_{X_i \circ h^\star}(t) - q_{X_j \circ h^\star}(t)\right)^2 dt = \\ &= \int_T \left(q_{X_i}(h^\star(t))\sqrt{\frac{dh^\star}{dt}(t)} - q_{X_j}(h^\star(t))\sqrt{\frac{dh^\star}{dt}(t)}\right)^2 dt = \\ &= \int_T \left(q_{X_i}(t) - q_{X_j}(t)\right)^2 dt = d^2_{ETW}(X_i, X_j) \end{aligned}$$

and symmetry

$$d_{ETW}(X_i, X_j \circ h) = d_{ETW}(X_i \circ h^{-1}, X_j)$$

in contrast to the standard L_2 norm used in SMTW or PCS.

Instead of alignment of the original curves, the main idea of ETW is the alignment of corresponding SRSFs to a target function called *Karcher mean* q_μ [Tucker et al., 2013]. The warping functions $h_1, \ldots, h_N$ are found by minimising the criterion

$$S_{ETW}(h_i) = \int_T \left(q_\mu(t) - (q_{X_i} \circ h_i)(t)\sqrt{h_i'(t)}\right)^2 dt, \;\; i = 1, \ldots, N \tag{4.9}$$

by the dynamic programming algorithm. More details about the algorithm can be found in [Tucker et al., 2013] or in the documentation to the R package *fdasrvf* [Tucker, 2016].

The ETW method is one of the most powerful methods for curves registration. On the other hand, because of no restrictions to the distance between the real time and the warping function, ETW may produce close to ideal alignment of possibly dissimilar curves. This problem will be demonstrated in Section 4.4. Fortunately, it is possible to add a penalty to the cost function (4.9), for example

$$S^\star_{ETW}(h_i) = S_{ETW}(h_i) + \lambda \int_T \left(1 - \sqrt{h_i'(t)}\right)^2, \;\; i = 1, \ldots, N. \tag{4.10}$$

Optimal value for the penalty parameter λ can be selected in a similar way as in the modified SMTW algorithm (Section 4.3.1). Other types of penalties may be found in [Srivastava and Klassen, 2016, Chapter 8].

4.4 Application to sleep data

In this section we applied the above mentioned registration methods to the sleep probabilistic curves of subjects without sleep problems from the SIESTA database. We would like to highlight that the analysis of the sleep structure in the functional sense and the time alignment of the sleep probabilistic curves are new approaches in the area of sleep research and we are not aware of any other scientific teams which would consider this approach.

The registration methods described in the previous sections require the assumption of the common start and end point. In our case, the subjects went to bed approximately at the same time, but the duration between the lights–off and falling asleep differed among subjects. Therefore the starting point for each sleep probabilistic curve was set to the sleep latency. Rechtschaffen and Kales [1968] define the sleep latency as three consecutive periods (of length 30 seconds) of the *S1* stage or the first appearance of the *S2* stage, whichever comes first.

To avoid noise the curves were smoothed by the FPCA and smoothing covariance surface [Yao et al., 2003]. An interesting by–product of the smoothing procedure is a reconstruction of the sleep probabilistic curves at the end of night according to behaviour of the whole dataset, therefore the smoothed curves are defined on the common time interval. For more details see [Yao et al., 2003].

In this section we focused on the time alignment of 146 sleep probabilistic curves representing sleep Microstate 1 (85% *S2*) of the second night. We considered the SMTW, PCS and ETW method for the registration. In the case of SMTW and ETW both, the original and penalised versions were used.

Because of many different sleep profiles present in our sleep curves dataset (Figure 4.5a), the direct registration of the whole dataset doesn't yield to reasonable results. Specifically,

- the PCS method aligns each pair of curves separately and then for each curve the global warping function is computed. However, because of the presence of many different sleep profiles, the global warping functions were close to real time and consequently the visual alignment of the sleep probabilistic curves was poor (Figure 4.5b).

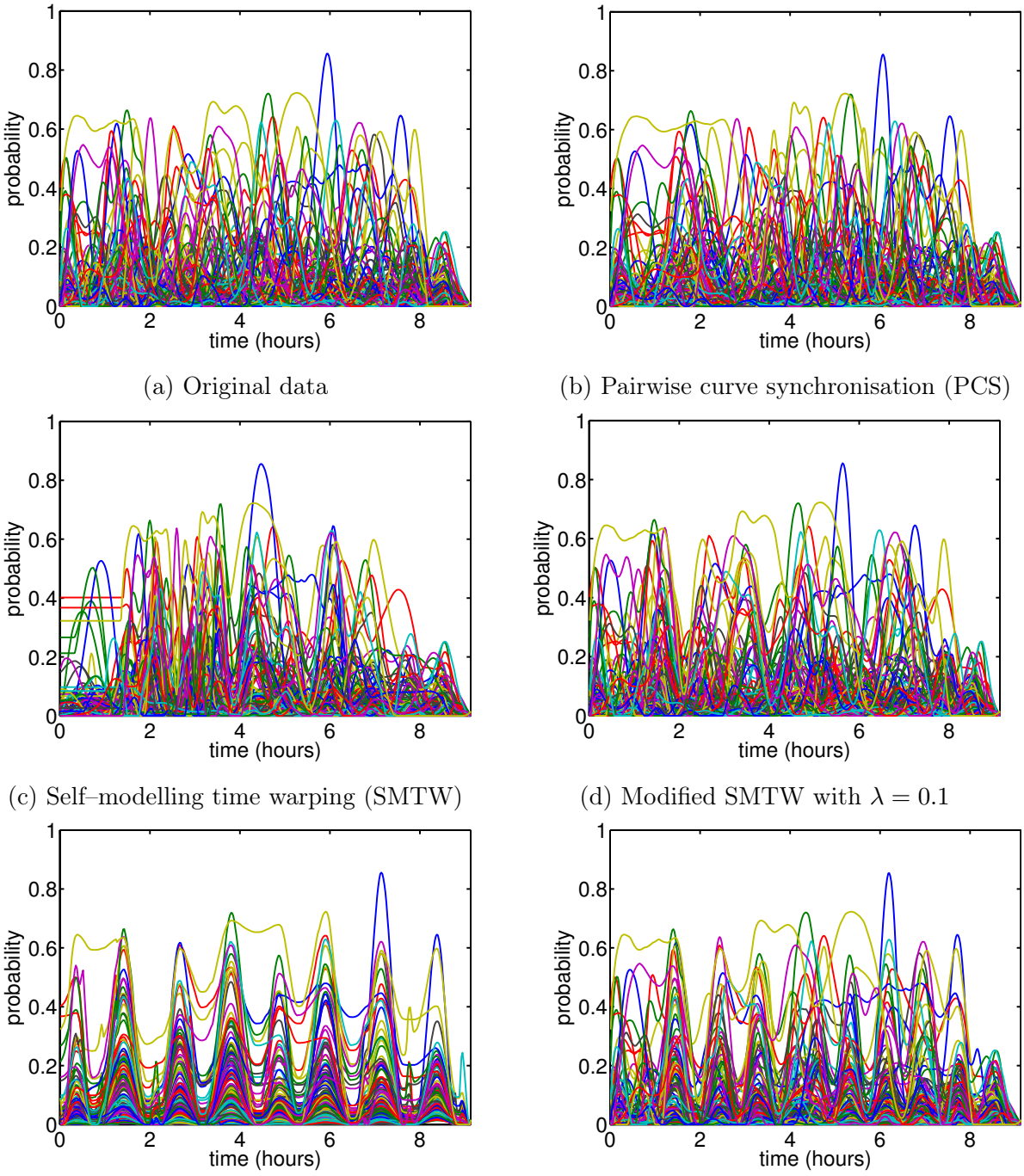
- considering the original SMTW algorithm, unexpected flat segments occurred in several curves right after the starting point (Figure 4.5c). These flat segments were

caused by improper penalisation of nondecreasing warping functions in SMTW (Figure 4.6) and due to the alignment of the curves with different profiles to one target. After applying the proposed penalised version of the SMTW algorithm, the flat segments diminished but at the cost of poor alignment of the curves (Figure 4.5d).

- the ETW method visually aligned the curves well (Figure 4.5e), but this is at the cost of rapid changes of the curves profiles (Figure 4.7). The ETW method is one of the most powerful methods for the curves alignment. However, when trying to align sleep curves with different profiles to one target, the important curve elements were shifted too far in time which may lead to misinterpretation of results. Considering the penalised version of ETW with $\lambda = 0.1$ the curves were visually aligned better than by PCS or SMTW, but the typical sleep profiles were difficult to detect.

Finally we can conclude that the sleep database includes too many different sleep profiles and therefore the direct registration within each sleep microstate separately results either in improper alignment or on the other hand in too ideal alignment of possibly dissimilar curves and rapid distortions of the curves shapes. Therefore it would be better to align only curves with similar shapes. However, this in turn leads us back to the cluster analysis of the misaligned curves.

These observation underline the need for the approach which combines clustering and curves alignment steps. In the next section we give an overview of existing methods for simultaneous curves alignment and clustering and appropriateness of their application to the analysed sleep data. In the centre of interest is our own method which is experimentally validated and compared with existing approaches.



(a) Original data

(b) Pairwise curve synchronisation (PCS)

(c) Self–modelling time warping (SMTW)

(d) Modified SMTW with $\lambda = 0.1$

(e) Elastic time warping (ETW)

(f) Penalised ETW with $\lambda = 0.1$

Figure 4.5: Time alignment of 146 sleep probabilistic curves representing sleep Microstate 1 (85% *S2*). The curves were aligned by three different methods operating on the whole dataset. In the case of the Self–modelling time warping (SMTW) and Elastic time warping (ETW) both penalised and non–penalised versions of the algorithms were considered.

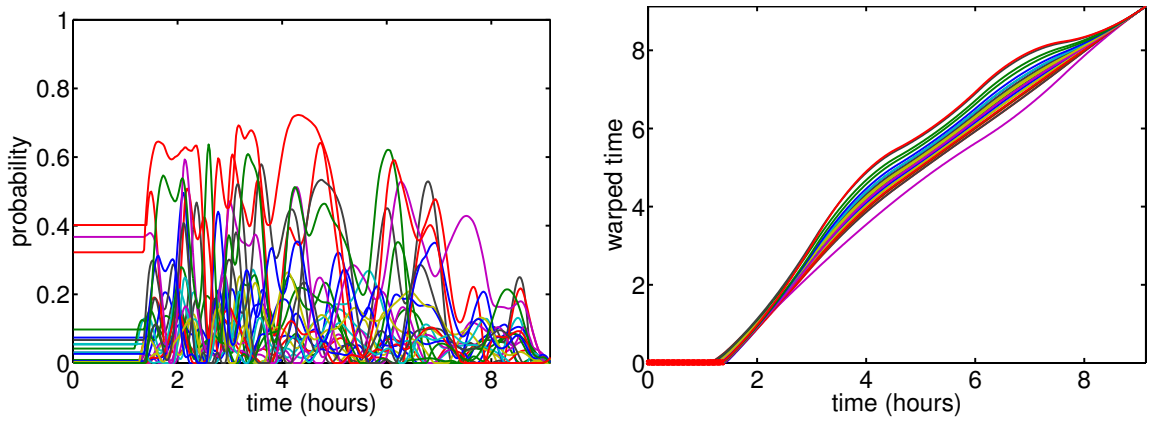


Figure 4.6: A subset of curves from Figure 4.5c aligned by the Self–modelling time warping algorithm (SMTW) (left) and corresponding warping functions (right). The constant segments in warping functions at the beginning of night are depicted in red.

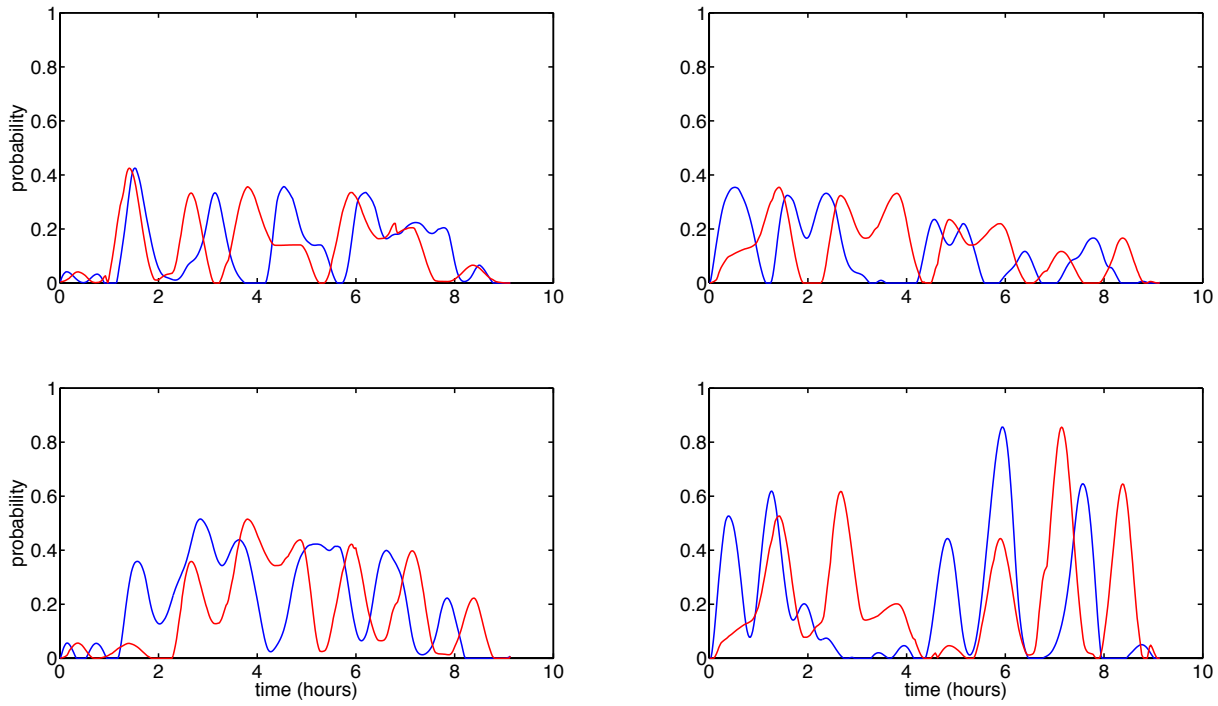


Figure 4.7: Example of four sleep probabilistic curves representing Microstate 1. The curves aligned by the ETW method (red) are shifted for more than one hour in comparison to their original version (blue).

Chapter 5

Methods for combination of the curves alignment and cluster analysis

In the previous chapter we described the idea of the curves alignment. Three chosen registration method were applied to the sleep probabilistic curves. However, because of many different curves profiles none of the methods produced satisfactory results. We observed that it is reasonable to first divide the curves with similar shapes and features into homogeneous subgroups and then to register curves in each subgroup separately. In other words, to apply cluster analysis before the registration process. But our initial goal was to register misaligned data before the clustering step. This is due to the fact that considering original misaligned curves may lead to improper assignment into clusters. It looks like we are facing "the chicken and the egg" problem.

Fortunately, there exist algorithms which combine cluster analysis and curves alignment. Mainly, these are

- k–mean alignment for curve clustering (KMACC) [Sangalli et al., 2010],
- Joint probabilistic curve clustering and alignment (JPCCA) [Gaffney and Smyth, 2005],
- the truncated version of PCS [Tang and Müller, 2009] which can be viewed as an algorithm which takes into account similarity between curves within the registration process. In the thesis the abbreviation tPCS denotes the curves alignment produced by the truncated PCS algorithm followed by the k–means clustering.

The first two methods KMACC and JPCCA operate with a linear transformation of

time

$$h(t) = at + b, \quad a > 0, \ b \in \mathbb{R}, \ t \in T$$

when solving the curves misalignment problem. However, this limits the flexibility of the methods to deal with situations where a nonlinear transformation of time is needed. In addition, we consider the same time interval for all sleep curves and therefore the only possible choices for the a and b constants are $a = 1$ and $b = 0$, effectively producing no alignment. Other values of a or b would annul the property of the common time interval.

The KMACC method considers either the L^2 distance (3.2) or the distance measure (3.3). The input are either original curves or their first derivatives. However, we miss a restriction between the warping function and the real time. When we relaxed the assumption of the common start and end point and tried to apply the KMACC algorithm to the sleep probabilistic curves we observed poor alignment or the situation where the curves were shifted for more than 4 hours and consequently they missed their physiological interpretation.

The JPCCA method is based on a functional regression mixture model. The correct cluster membership is considered as a latent variable, therefore the *EM* algorithm is used to estimate unknown parameters in the model. However, when applied to our sleep data the *EM* algorithm sometimes failed to converge due to the unstable estimation of covariance matrices or the number of clusters collapsed to one.

The only algorithm combining the curves alignment and clustering which seems to be appropriate for our sleep data is tPCS. However, the results produced by tPCS were usually not satisfactory (Section 6.1). Therefore we designed our own 2–step approach which consists of the Dynamic time warping based clustering and an arbitrary registration method. This represents one of the important new contributions of this PhD thesis.

5.1 A 2–step approach for functional data clustering

To address the problems of the existing methods for combined clustering and registration and also to introduce an algorithm with a higher flexibility of algorithmic choices in the registration step, we propose a new 2–step approach.

We start with introducing a clustering method based on the Dynamic time warping method [Wang and Gasser, 1997].

5.1.1 Dynamic time warping as a clustering step

The DTW algorithm is a method which was a priori developed and used for aligning curves with different lengths [Parsons, 1987]. In contrast to the other registration methods the algorithm works directly with a discrete representation of curves.

Let suppose that two curves X_1, X_2 defined on the time interval T are observed at a finite number of time-points

$$x_1 = \{X_1(t_1), \ldots, X_1(t_{n_1}), 0 = t_1 < \cdots < t_{n_1} = 1\},$$
$$x_2 = \{X_2(s_1), \ldots, X_2(s_{n_2}), 0 = s_1 < \cdots < s_{n_2} = 1\}.$$

It is not required that the sets of time-points $\{t_i\}_{i=1}^{n_1} \subset T$ and $\{s_i\}_{i=1}^{n_2} \subset T$ are equal.

The main goal of the DTW method is to find the best match between curves X_1 and X_2 by constructing the *warping path* $w = \{(i_l, j_l), i_l \in \{1, \ldots, n_1\}, j_l \in \{1, \ldots, n_2\}, l = 1, \ldots, W_L\}$, which minimises the cost function

$$S_{DTW}(X_1, X_2, w) = \sum_{(i_l, j_l) \in w} |X_1(t_{i_l}) - X_2(s_{j_l})|, \tag{5.1}$$

where W_L is the length of the warping path w. This problem can be solved by using the dynamic programming [Wang and Gasser, 1997].

In Section 4.2 we described several constraints on the warping function. With small modifications the constraints are incorporated also in the DTW method, namely

- $i_{l-1} \leq i_l \quad$ and $\quad j_{l-1} \leq j_l, \quad l \in \{1, \ldots, W_L\}$ (monotonicity),
- $i_1 = 1,\ j_1 = 1,\ i_{W_L} = n_1,\ j_{W_L} = n_2$ (common starting and end point),
- $i_l - i_{l-1} \leq 1,\ j_l - j_{l-1} \leq 1$ (continuity),
- $|i_l - j_l| \leq r, \quad r \in \mathbb{N}$ is a chosen constant which controls the distance between the warping path and the real time. In the sleep data analysis we vary r between 30 and 50. We observed that this provides good performance.

In the simplest form the DTW algorithm is able to align only a pair of curves. Nevertheless, it results in an interesting by–product which is the distance measure [1] defined in (5.1) and which we denote as dtw

$$dtw(X_1, X_2) = \min_w \sum_{(i_l, j_l) \in w} |X_1(t_{i_l}) - X_2(t_{j_l})|.$$

[1] We use the term "distance" despite the formula (5.1) does not show the symmetry property and therefore it is not a real distance.

In our thesis we use dtw for constructing a matrix $M_{dtw} \in \mathbb{R}^{N \times N}$

$$(M_{dtw})_{ij} = dtw(X_i, X_j), \quad i, j = 1, \ldots, N. \tag{5.2}$$

The matrix M_{dtw} represents a distance matrix which can be used in the hierarchical or k–medoids functional data clustering algorithms [Montero and Vilar, 2014].

5.1.2 2–step approach

The general idea of the 2–step approach is based on a direct combination of the clustering and registration steps into an iterative process, more specifically

i) In the first step, an initial clustering is done. Because at this first step the curves are misaligned, the standard k–means or k–medoids algorithms does not lead to reasonable results. Therefore, in this initial clustering step, we propose to apply the DTW method with the aim to obtain the distance matrix M_{dtw} and then to apply the k–medoids clustering algorithm operating on M_{dtw}.

ii) In the second step, we register misaligned curves separately in each cluster. In practice, one of the above mentioned three algorithms (SMTW, PCS or ETW) or their penalised versions can be used. The quality of alignment and clustering is measured by the L–criterion (3.4).

iii) The third step consists of re–clustering of the aligned curves using the same clustering approach as in the step i).

Steps ii) and iii) are the core steps of the 2–step approach for iterative clustering and alignment and are repeated until one of the following stoping criteria is met

- the number of iterations exceeds a given threshold (in this thesis set to 100),
- the L–criterion is lower than a given small constant,
- clusters in the $i^{th}, (i-1)^{th}$ and $(i-2)^{th}$ steps are not changed.

The chosen stopping criteria mimic those used in the standard clustering techniques, for example [Lloyd, 1982; Kaufman and Rousseeuw, 1990]. Finally, the cluster membership and aligned curves belonging to the iteration step with the smallest L–criterion are used as the final result.

From a mathematical point of view the 2–step approach represents an heuristic and not a fully rigorous mathematical method. However, our long–term analysis of the sleep data showed, that the sleep probabilistic curves are too difficult data for many existing methods with a more complex mathematical background. These existing methods were outperformed by the proposed approach.

Chapter 6

The time alignment and cluster analysis of the sleep probabilistic curves

In the functional data analysis the registration of in time misaligned curves is a standard preprocessing step preceding further analysis [Ramsay and Silverman, 2005]. However, when discussing the time alignment procedure of the sleep probabilistic curves with authorities in the area of sleep research, two important questions occurred. First, don't we partially miss information about the sleep structure after the time alignment of the curves? And if yes, for which sleep microstates the exact occurrence of periods of high probabilities is important when detecting relationships between them and daily measures?

In this chapter we aimed to find answers to these questions by applying the 2–step approach separately to the database of healthy sleepers and patients after stroke.

6.1 Cluster analysis of the sleep structure of healthy sleepers

In order to demonstrate the benefit of the curves alignment we choose the tPCS algorithm and the 2–step approach as representatives for the approaches combining the time alignment and the cluster analysis of the sleep probabilistic curves. The k–means clustering of the raw sleep curves is also considered in this thesis and serves as a reference allowing us to compare the obtained results with the clustering operating on in time misaligned curves.

In the first step we performed the analysis of the sleep data from the SIESTA database. To avoid possible misinterpretation of the results due to the first night effect, in this thesis

only data recorded during the second night are considered.

The cluster analysis of the microstates is divided into five subsections according to the similarity to the *slow wave sleep (SWS)*, *REM* stage, *S2* stage and finally wakefulness or light sleep. Regarding to the 2–step approach we considered its version with the modified SMTW method, PCS and penalised version of the ETW algorithm used in the registration step; that is 2DTW–SMTW, 2DTW–PCS, 2DTW–ETW.

Average silhouette	number of clusters	k–means	tPCS	2DTW– –SMTW	2DTW– –PCS	2DTW– –ETW
Microstate 16	8	0.56	0.57	0.64	0.61	0.47
Microstate 8	9	0.33	0.34	0.48	0.44	0.46
Microstate 14	3	0.53	0.54	0.60	0.10	0.60
Microstate 1	2	0.77	0.79	0.79	0.64	0.80
Microstate 6	3	0.62	0.63	0.73	0.71	0.53
L-criterion	number of clusters	k–means	tPCS	2DTW– –SMTW	2DTW– –PCS	2DTW– –ETW
Microstate 16	8	0.40	0.39	0.26	0.33	0.19
Microstate 8	9	0.79	0.77	0.54	0.59	0.34
Microstate 14	3	4.55	4.51	2.71	5.01	1.77
Microstate 1	2	4.07	4.05	3.37	4.34	2.84
Microstate 6	3	1.02	1.01	0.78	0.85	0.73

Table 6.1: The average silhouette and the L–criterion eq. (3.4) values for methods used to validate the alignment and clustering performance on 146 probabilistic sleep curves of several microstates. The optimal number of clusters for each microstate is depicted in the second column.

Microstates similar to *SWS*

Microstate 16 represents *SWS* with probability 96%. Using the L–criterion and average silhouette (AS) the optimal number of clusters was set to 8.

The k–means clustering of misaligned curves and the tPCS algorithm produced similar results (Figure 6.1). The tPCS algorithm was not able to properly align curves and the pattern typical for *SWS* is difficult to detect. Using the 2–step approach, visual

improvement in alignment of curves is depicted in Figure 6.2. A significant decrement in the L–criterion together with the AS increment (Table 6.1) was also observed.

By considering the SMTW, PCS or ETW algorithm for registration in the 2–step approach within each cluster separately, clusters of different structure were created. The 2DTW–SMTW and 2DTW–PCS approaches detected one big cluster including curves with low probability values (cluster 7 in Figure 6.2a or cluster 4 in Figure 6.3). On the other hand, using the ETW method with $\lambda = 0$, the 2–step approach was able to distinguish three different clusters with low probability values (clusters 5, 6, 7 in Figure 6.2b). However, in Section 6.1.1 we will show, that all three clustering methods produced similar relationships between the structure of Microstate 16 and daily measures.

Finally we observed, that for Microstate 16 all five approaches formed one cluster with possibly an outlier profile (Figure 6.1 and 6.2).

Microstates similar to *REM*

Microstates 8 and 14 are similar to the *REM* stage with probabilities 74% and 72% respectively. A clear periodic pattern is typical for the sleep probabilistic curves of both sleep microstates. Using the L–criterion and AS, the optimal number of clusters for Microstate 8 was set to 9 and for Microstate 14 to three.

Considering the k–means clustering of misaligned curves or tPCS, the periodic pattern in the formed clusters is difficult to detect (Figures 6.4a and 6.5a). Moreover, for Microstate 8 both algorithms labeled one sleep profile as a possible outlier (cluster 1 in Figure 6.4a). However, from the visual point of view the profile is not very different from the shapes of the curves assigned to cluster 3 in Figure 6.4a.

A significant increment in the AS together with a decrement in the L–criterion was observed when using the 2–step approach in comparison to tPCS or the k–means clustering (Table 6.1). The three versions of the 2–step algorithm (2DTW–SMTW, 2DTW–PCS, 2DTW–ETW) produced equivalent results considering the AS values as well as visual inspection of the curves in the formed clusters (Figure 6.4b).

In contrast, in the case of Microstate 14, the PCS algorithm used in the registration step within the 2–step approach was not able to align the curves properly. An visual improvement in alignment was observed only after considering the 2–step approach with the SMTW or ETW method with $\lambda = 0.02$ (Figure 6.5b). Both versions of the 2–step algorithm achieved approximately the same values of the AS and L–criterion.

Microstates similar to *S2*

In this section we provide a detailed analysis of sleep Microstate 1 (85% *S2*). For all five investigated methods, the most visible drop in the L–criterion was present for two clusters. This choice of two clusters was also confirmed by the AS values.

The k–means algorithm applied to the misaligned curves (Figure 6.6a) divided the whole set into two parts – curves with probability values of Microstate 1 under 0.4 and curves with higher probability values. However, the typical sleep profile of each cluster is difficult to detect.

Visually observed similarity between results obtained by the k–means clustering of the misaligned curves and the curves aligned by tPCS before the k–means clustering is mirrored also in similar values of the L–criterion and AS (Table 6.1).

The lowest value of the L–criterion together with the highest AS was reached when using the 2DTW–SMTW version of the 2–step algorithm (Figure 6.6b). The data assigned into the first cluster are visually aligned to three or four dominant peaks and some level of alignment can also be noticed in the second cluster. The L–criterion decreased and a higher AS value was observed for 2DTW–SMTW (Table 6.1) in comparison to the tPCS or the k–means clustering of misaligned curves.

Microstates similar to *Wake*

The PSM distinguishes four sleep microstates related to the *Wake* stage or light sleep. In this section we analyse only the most specific one – Microstate 6 (85% *Wake*) which represents the periods of full awakening during and at the end of the night (Figure 6.7a). The optimal number of clusters was chosen to be three.

The first two approaches – the k–means clustering of misaligned curves and tPCS – formed the clusters according to the position and length of the awake time interval at the end of the night. On the other hand, the 2–step approach with SMTW or PCS distinguishes the first cluster with subjects being approximately one hour awake before the final awakening, the second cluster with lower probability values for awakening during the whole night and the third cluster with two possibly outlier profiles (Figure 6.7b). We would like to highlight, that because of the alignment, the exact length of the time periods of wakefulness is not taken into account in the process of forming the clusters. Influence of this phenomenon on the relationships with daily measures will be discussed in Section 6.1.1.

When considering the ETW algorithm with $\lambda = 0$ (2DTW–ETW), rapid distortion

of the curve profiles is visible (Figure 6.7c). Moreover, the AS is lower in comparison to the 2DTW–SMTW or 2DTW–PCS methods, but also in comparison to the k–means clustering of misaligned curves or tPCS.

Standard sleep stages

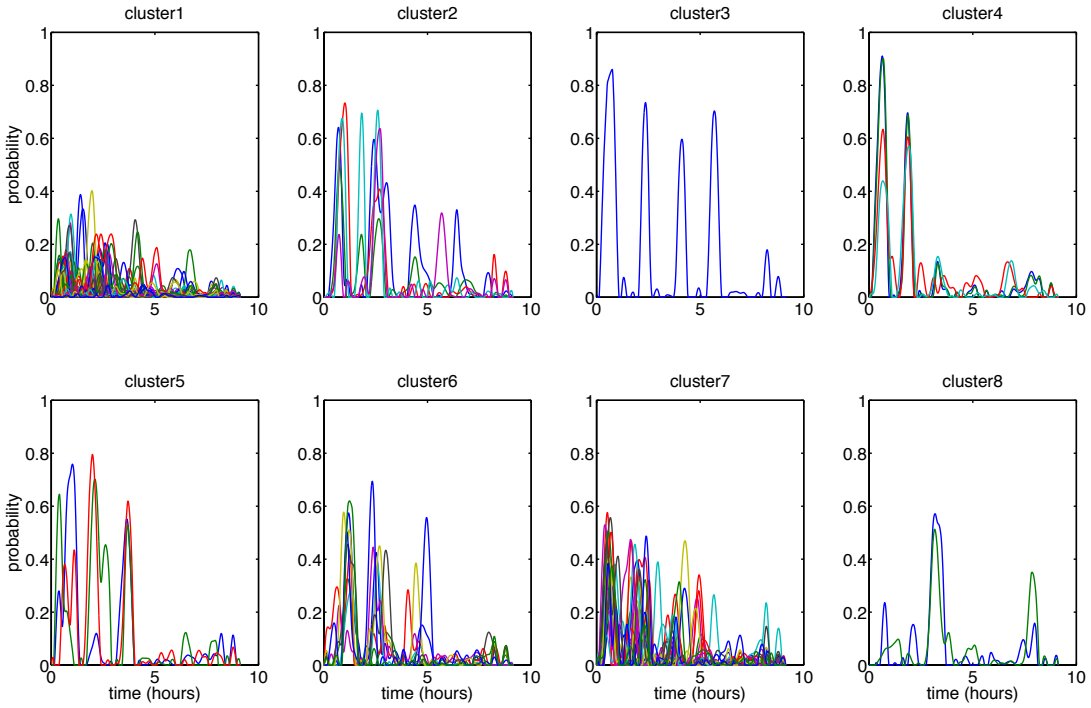
The same procedure was applied also to sleep probabilistic curves of the standard sleep stages *Wake, S1, S2, SWS* and *REM*. In all five cases the optimal number of clusters was set to two.

Using the k–means clustering of in time misaligned curves for the *Wake* stage, the representatives of clusters are difficult to detect (Figure 6.8). On the other hand the 2DTW–SMTW algorithm divided subjects into clusters according to the position of a set of peaks with higher probability for wakefulness during the night. For subjects assigned into cluster 2 in Figure 6.8b two periods of wakefulness are typical – approximately in the middle of the night and at the end of the night. A common feature of the subjects from cluster 1 are two periods of wakefulness located at the end of the night. Occurrence of short intervals of increased probability for the *Wake* stage in the first half of the night represents individuality of subjects from cluster 1.

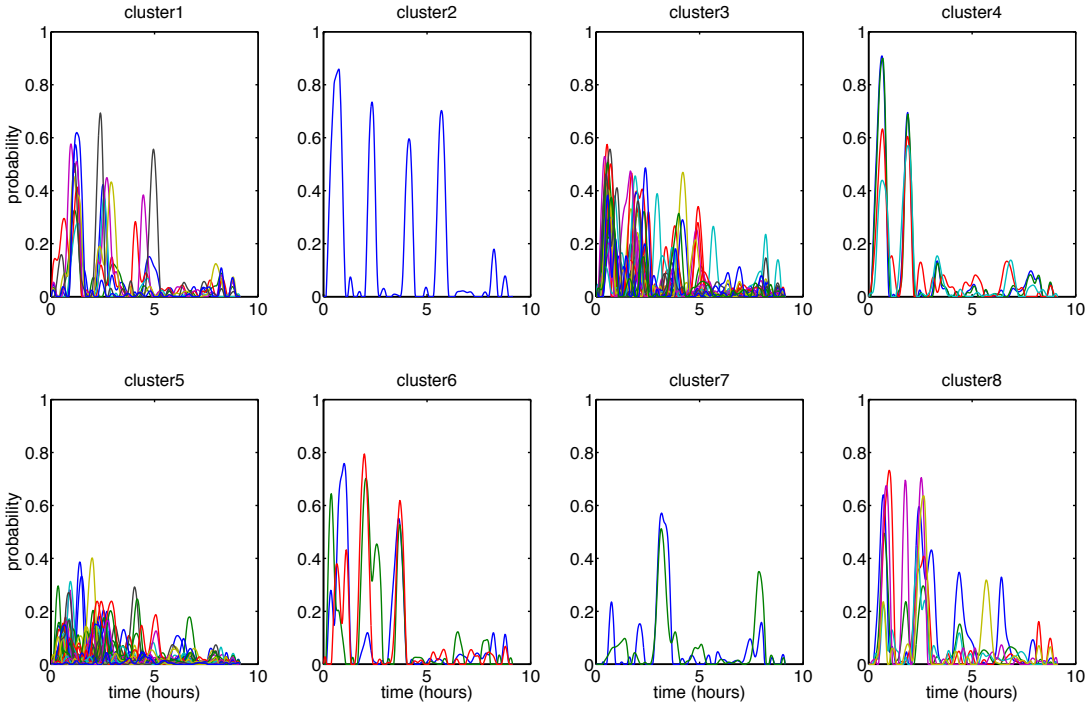
Sleep probabilistic curves for the *S1* stage are upper bounded by the value 0.4. All five considered methods led to similar structure of clusters (Figure 6.9). The tPCS, 2DTW–SMTW and 2DTW–PCS algorithms were not able to align the curves properly and visually they produced the same results as the k–means clustering of misaligned curves. Some kind of alignment was observed only in the case of 2DTW–ETW with $\lambda = 0.06$ (Figure 6.9b). In both clusters a periodic pattern and amplitude increasing with time is visible.

The *S2* stage is the major part of sleep spent by humans. This is reflected in the pattern of the sleep probabilistic curves with local maxima around 0.8. Similarly to the *S1* stage, the two clusters formed by any of the five methods differed especially in the amplitude (Figure 6.10a). Considering the curves alignment produced by 2DTW–SMTW (Figure 6.10b) a periodic pattern in both clusters was observed.

The most visible benefit of the curves alignment can be observed in the *SWS* and *REM* stage. For both sleep stages a periodic pattern is typical. In the case of the *SWS* the amplitude of peaks decreases with time, while for the *REM* stage the amplitude increases. This pattern is present in clusters formed by 2DTW–SMTW (Figures 6.11b and 6.12b), but it is difficult to detect in clusters formed from misaligned curves (Figures 6.11a and 6.12a).

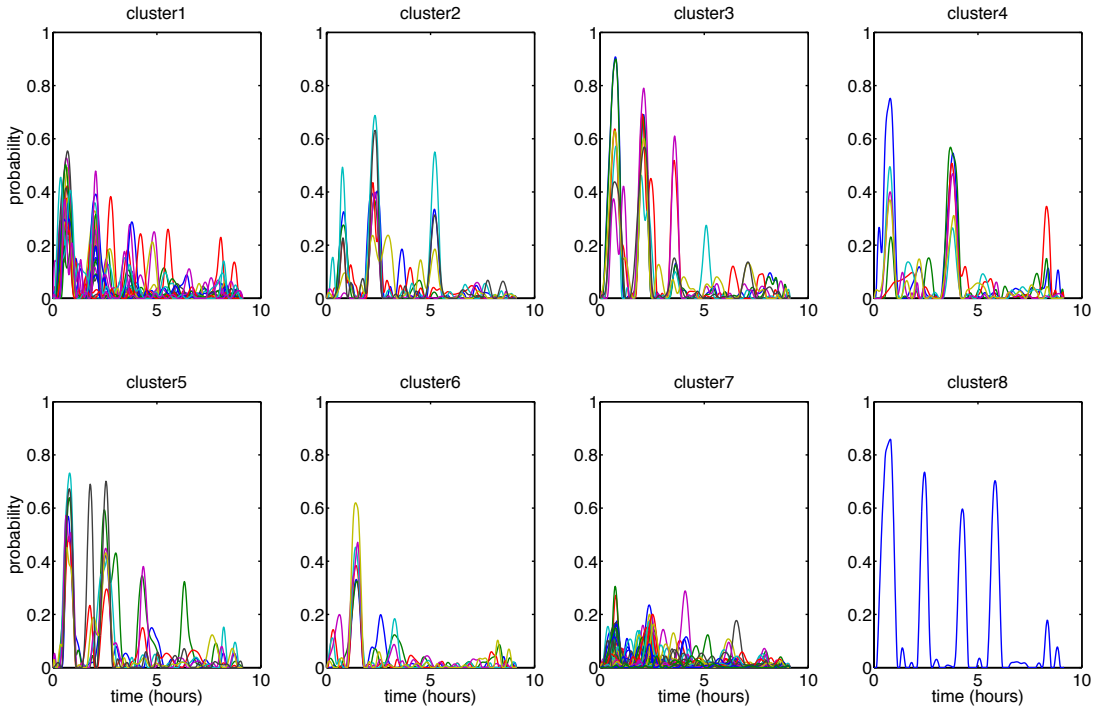


(a) The k–means clustering of misaligned sleep probabilistic curves

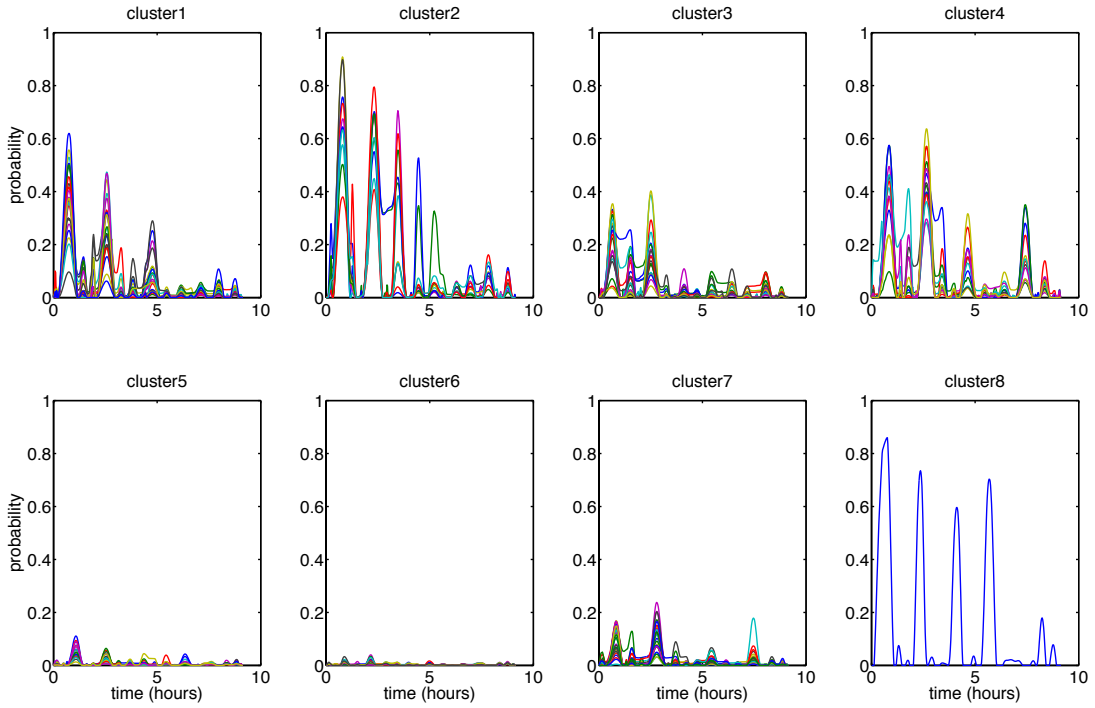


(b) tPCS followed by the k–means clustering

Figure 6.1: Microstate 16. Clustering of 146 probabilistic sleep curves into 8 clusters by using a) the k–means algorithm applied to misaligned curves and b) the truncated version of the Pairwise curve synchronisation algorithm (tPCS) followed by the k–means clustering.



(a) The 2–step approach with the modified SMTW algorithm (2DTW–SMTW)



(b) The 2–step approach with the ETW algorithm (2DTW–ETW)

Figure 6.2: Microstate 16. Clustering of 146 probabilistic sleep curves into 8 clusters by using the 2–step approach with the modified SMTW or ETW algorithm used in the registration step and k–medoids in the clustering step (2DTW–SMTW, 2DTW–ETW).

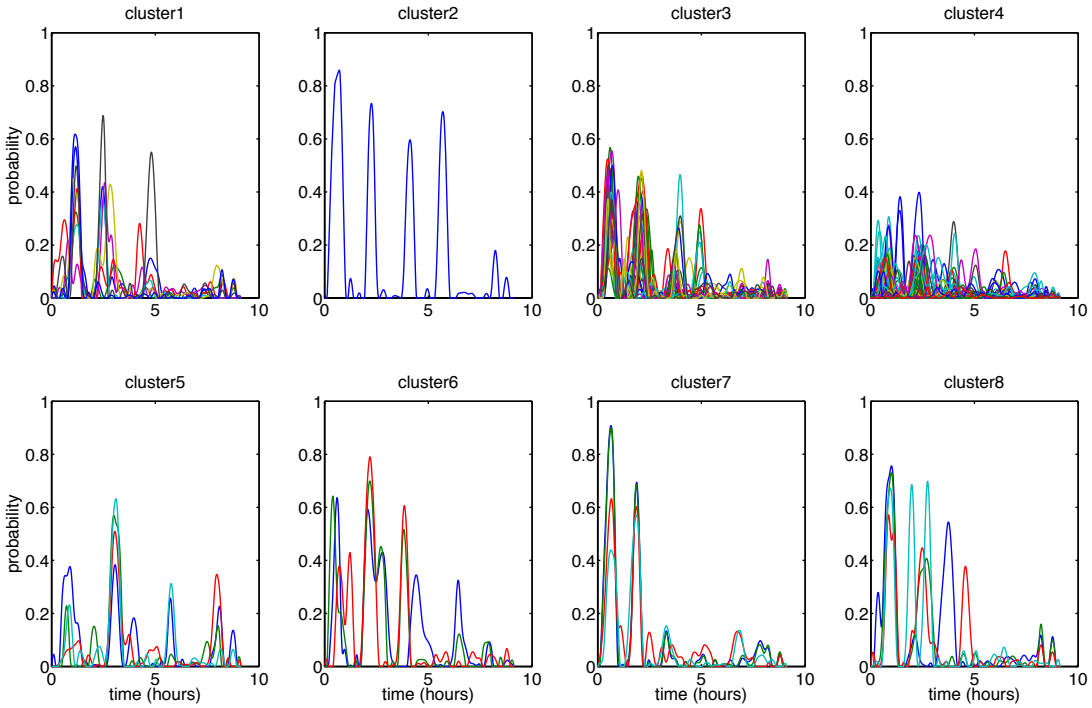
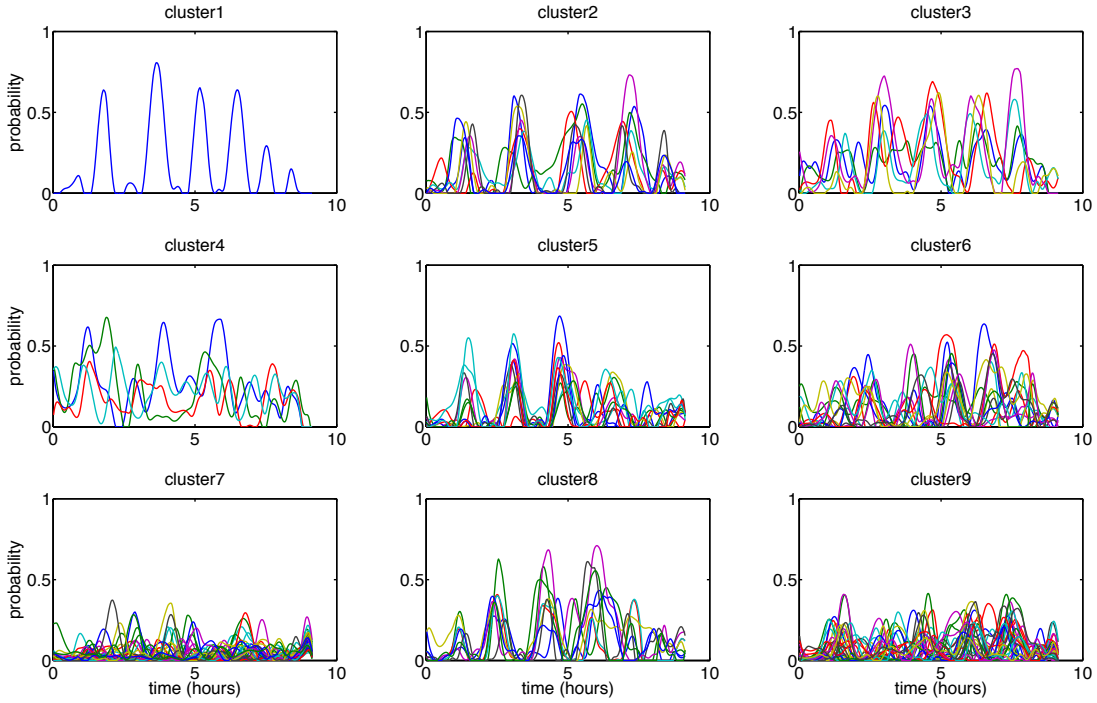
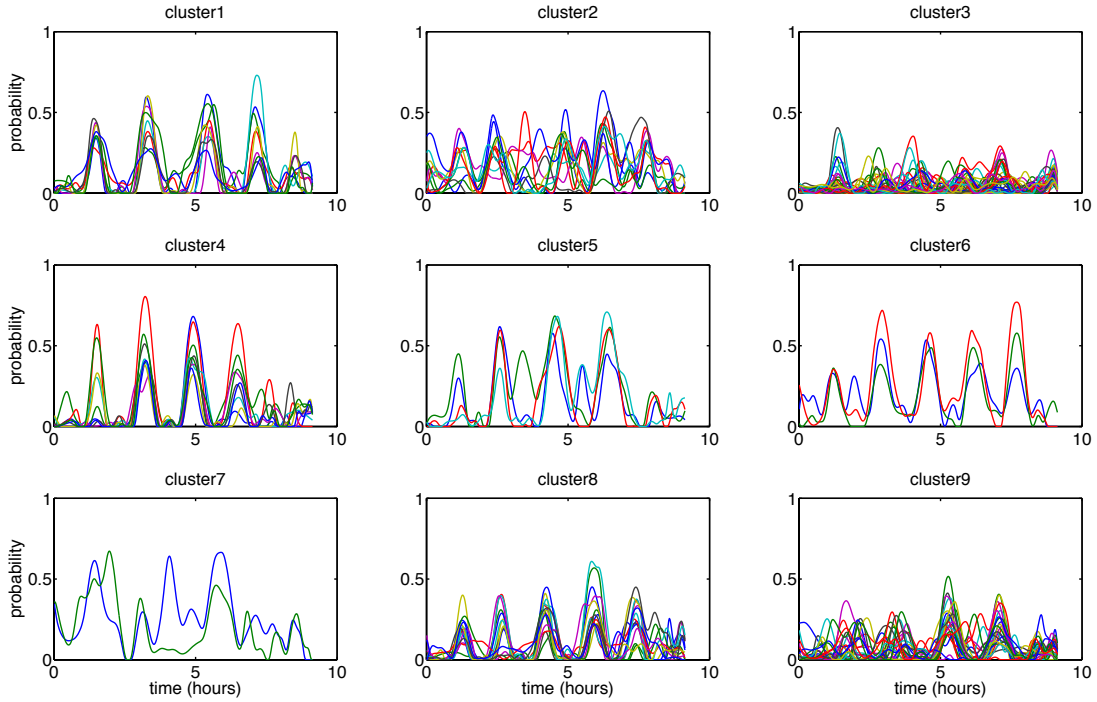


Figure 6.3: Microstate 16. Clustering of 146 probabilistic sleep curves into 8 clusters by using the 2–step approach with the PCS algorithm used in the registration step and k–medoids in the clustering step (2DTW–PCS).

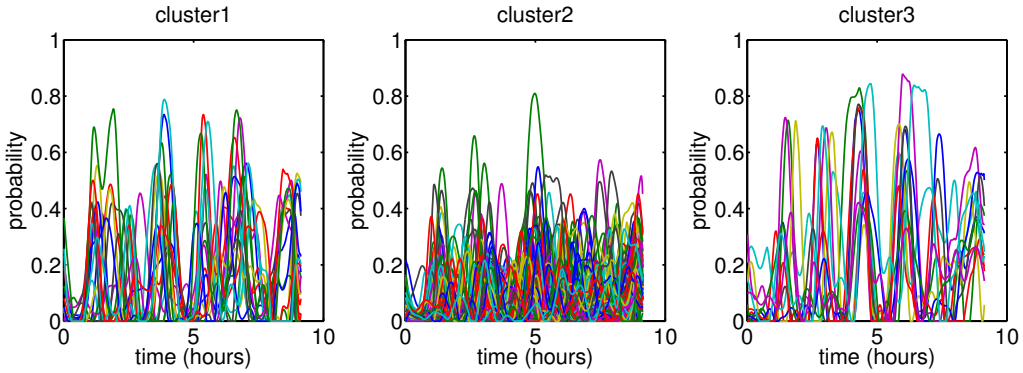


(a) The k–means clustering of misaligned sleep probabilistic curves

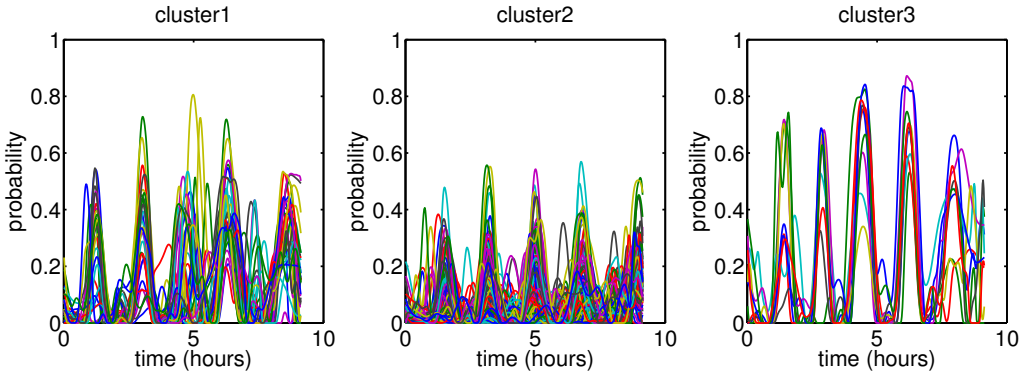


(b) The 2–step approach with the modified SMTW algorithm (2DTW–SMTW)

Figure 6.4: Microstate 8. Clustering of 146 probabilistic sleep curves into 9 clusters by using a) the k–means algorithm applied to misaligned curves and b) the 2–step approach with the modified SMTW algorithm in the registration step and k–medoids in the clustering step (2DTW–SMTW).

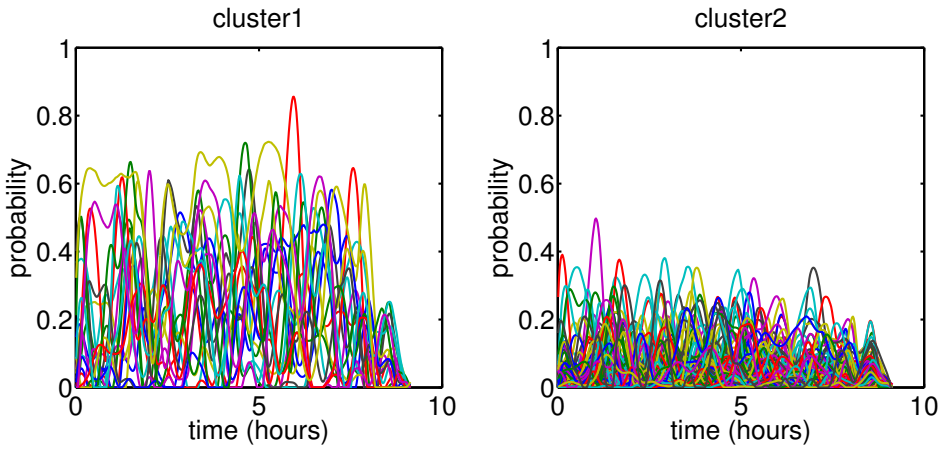


(a) The k–means clustering of misaligned sleep probabilistic curves

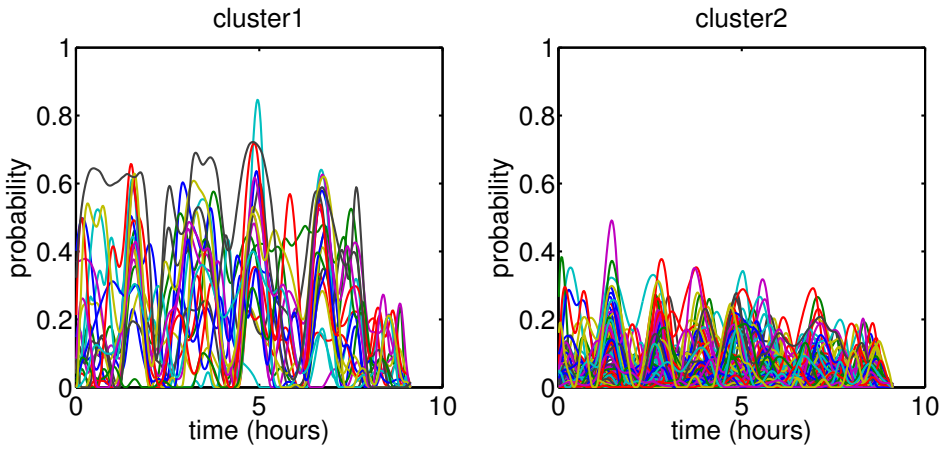


(b) The 2–step approach with the modified SMTW algorithm (2DTW–SMTW)

Figure 6.5: Microstate 14. Clustering of 146 probabilistic sleep curves into thee clusters by using a) the k–means algorithm applied to misaligned curves and b) the 2–step approach with the modified SMTW algorithm in the registration step and k–medoids in the clustering step (2DTW–SMTW).

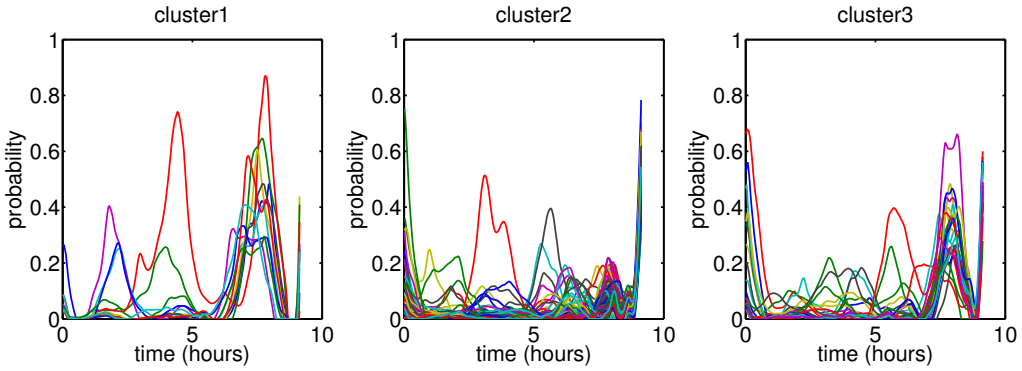


(a) The k–means clustering of misaligned sleep probabilistic curves

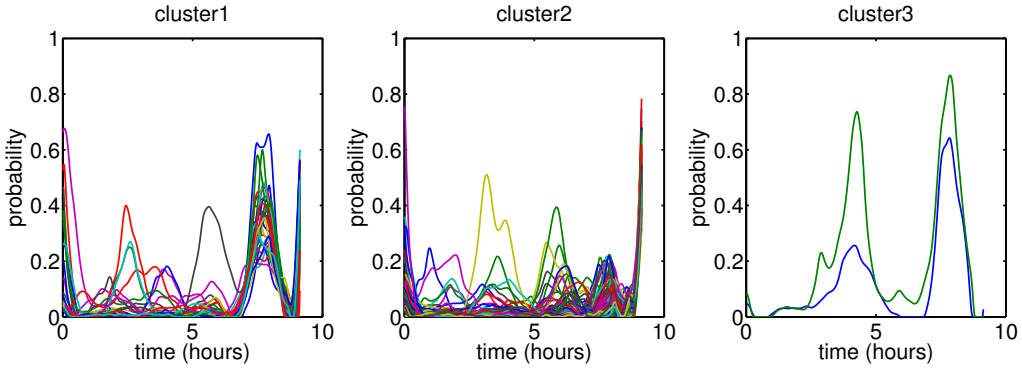


(b) The 2–step approach with the modified SMTW algorithm (2DTW–SMTW)

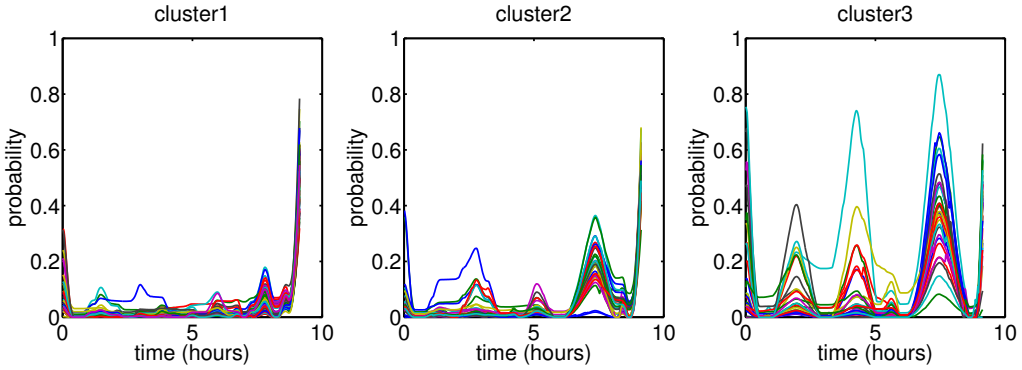
Figure 6.6: Microstate 1. Clustering of 146 probabilistic sleep curves into two clusters by using a) the k–means algorithm applied to misaligned curves and b) the 2–step approach with the modified SMTW algorithm in the registration step and k–medoids in the clustering step (2DTW–SMTW).



(a) The k–means clustering of misaligned sleep probabilistic curves

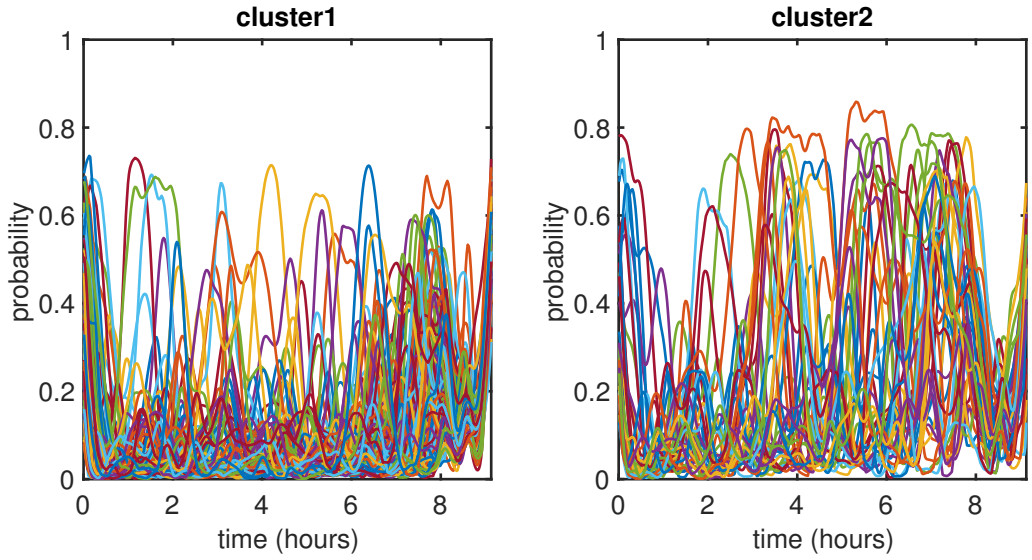


(b) The 2–step approach with the modified SMTW algorithm (2DTW–SMTW)

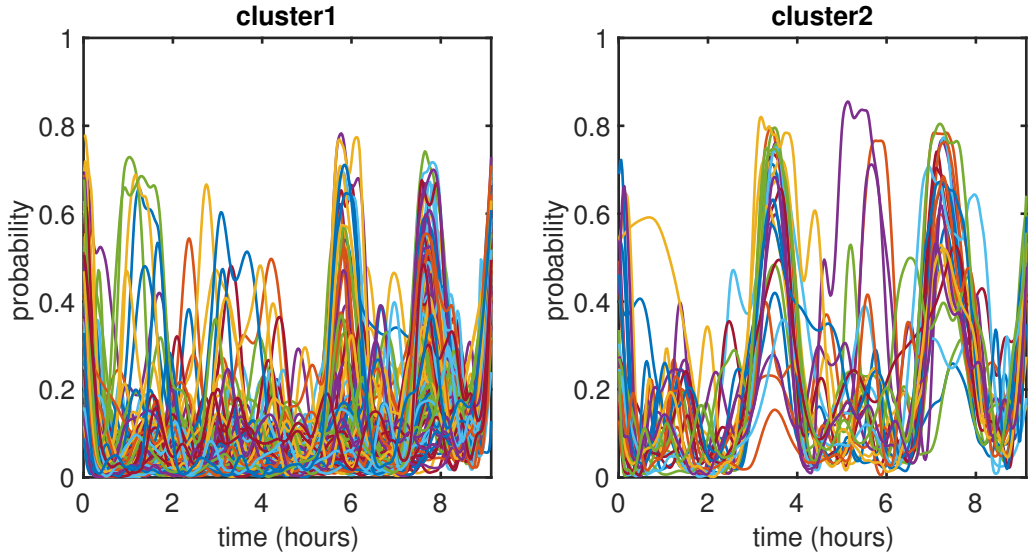


(c) The 2–step approach with the penalised ETW algorithm (2DTW–ETW)

Figure 6.7: Microstate 6. Clustering of 146 probabilistic sleep curves into three clusters by a) the k–means algorithm applied to misaligned curves and the 2–step approach with b) the modified SMTW algorithm or c) penalised ETW algorithm in the registration step and k–medoids in the clustering step (2DTW–SMTW, 2DTW–ETW).

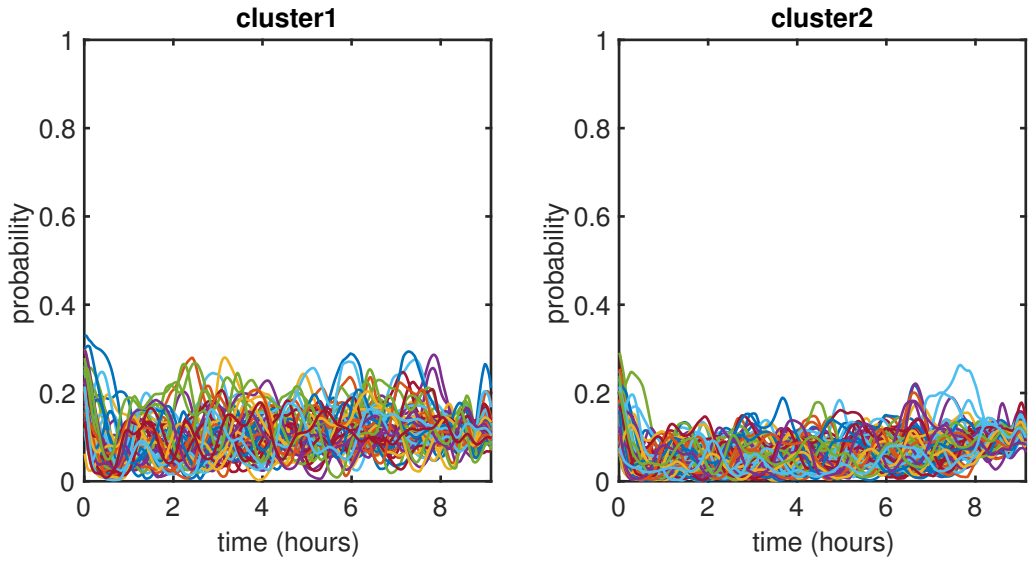


(a) The k–means clustering of misaligned sleep probabilistic curves

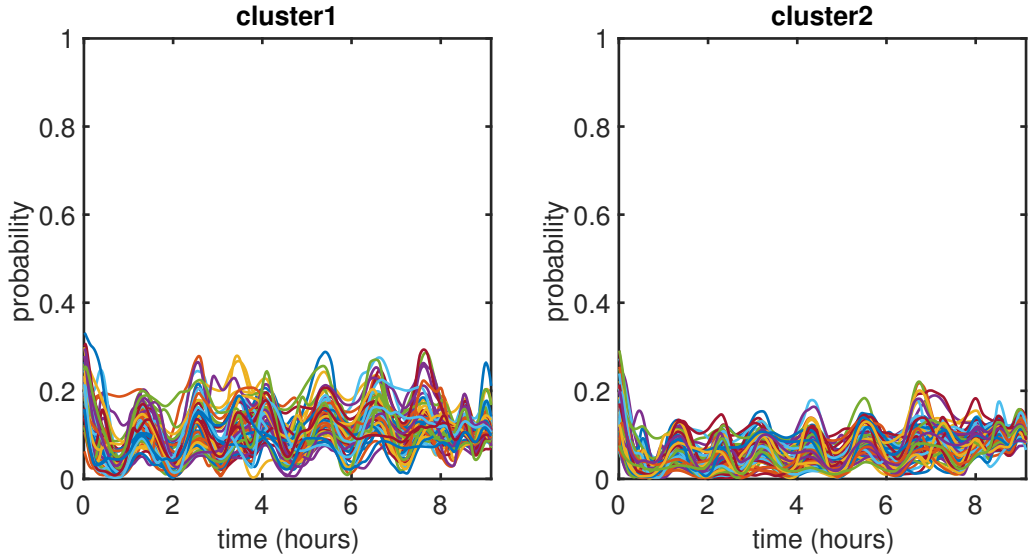


(b) The 2–step approach with the SMTW algorithm (2DTW–SMTW)

Figure 6.8: *Wake* stage. Clustering of 146 probabilistic sleep curves into two clusters by using a) the k–means clustering of misaligned curves and b) the 2–step approach with the modified SMTW algorithm in the registration step and k–medoids in the clustering step (2DTW–SMTW).

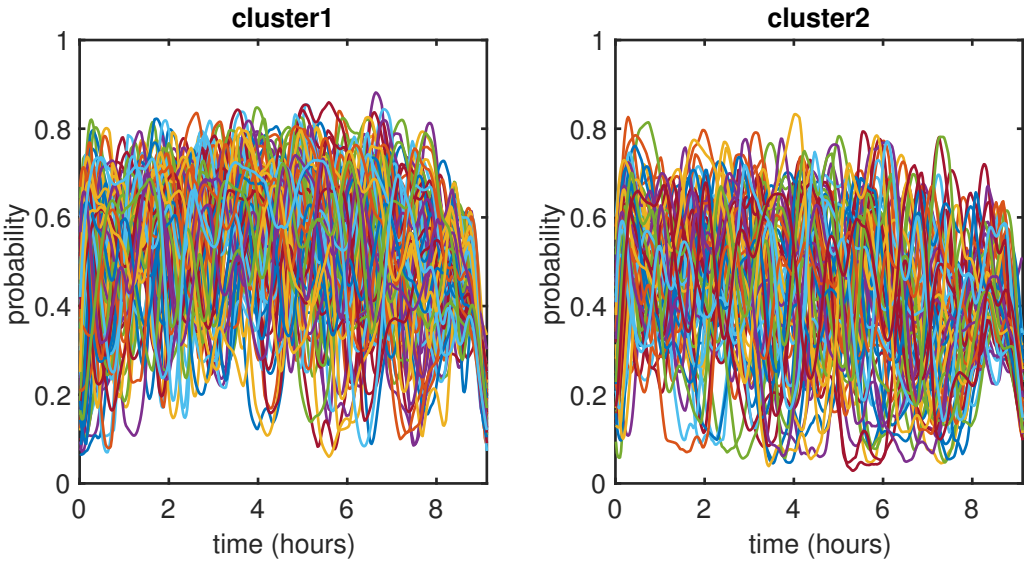


(a) The k–means clustering of misaligned sleep probabilistic curves

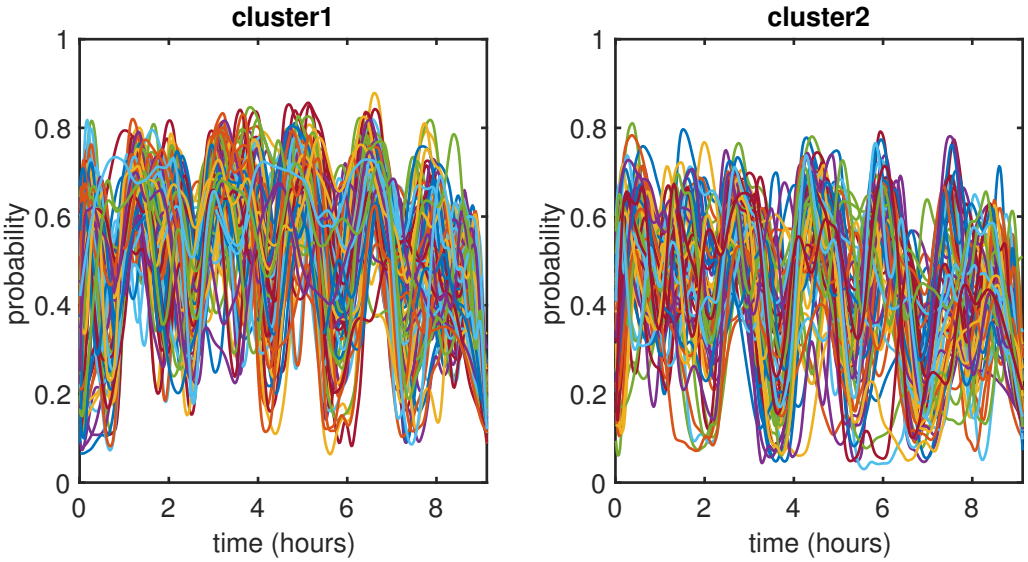


(b) The 2–step approach with the ETW algorithm (2DTW–ETW)

Figure 6.9: *S1* stage. Clustering of 146 probabilistic sleep curves into two clusters by using a) the k–means clustering of misaligned curves and b) the 2–step approach with the penalised ETW algorithm in the registration step and k–medoids in the clustering step (2DTW–ETW).

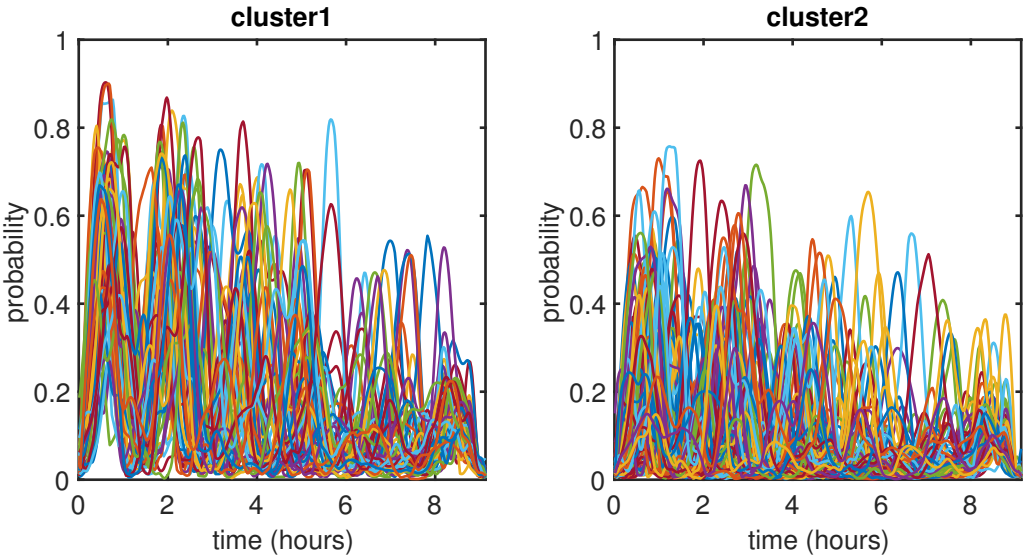


(a) The k–means clustering of misaligned sleep probabilistic curves

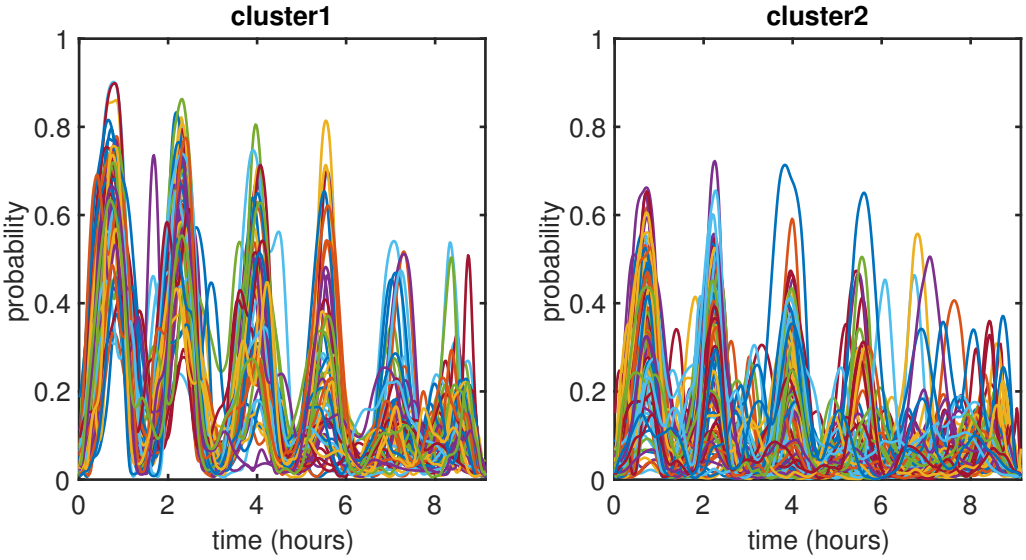


(b) The 2–step approach with the SMTW algorithm (2DTW–SMTW)

Figure 6.10: *S2* stage. Clustering of 146 probabilistic sleep curves into two clusters by using a) the k–means clustering of misaligned curves and b) the 2–step approach with the modified SMTW algorithm in the registration step and k–medoids in the clustering step(2DTW–SMTW).

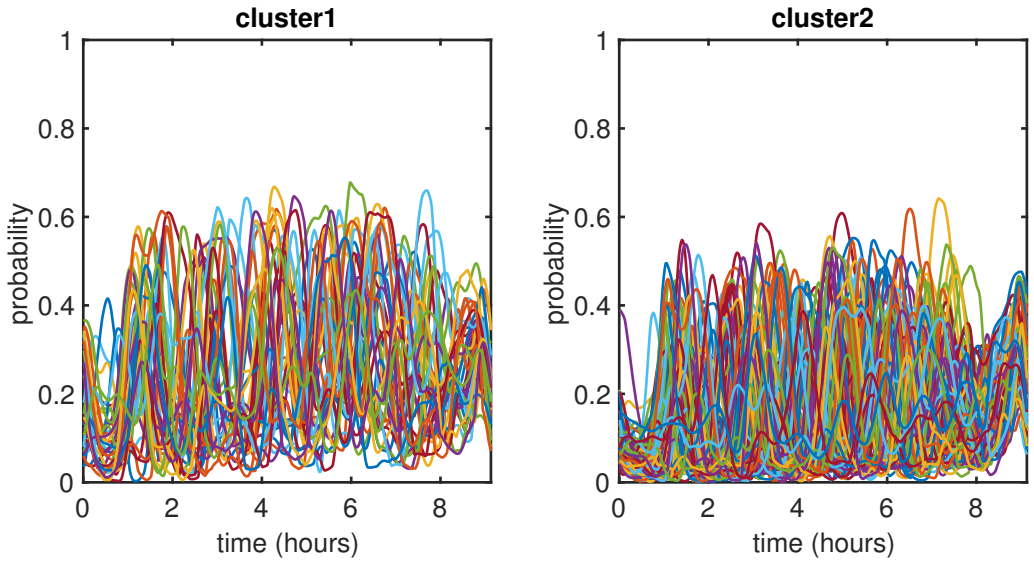


(a) The k–means clustering of misaligned sleep probabilistic curves

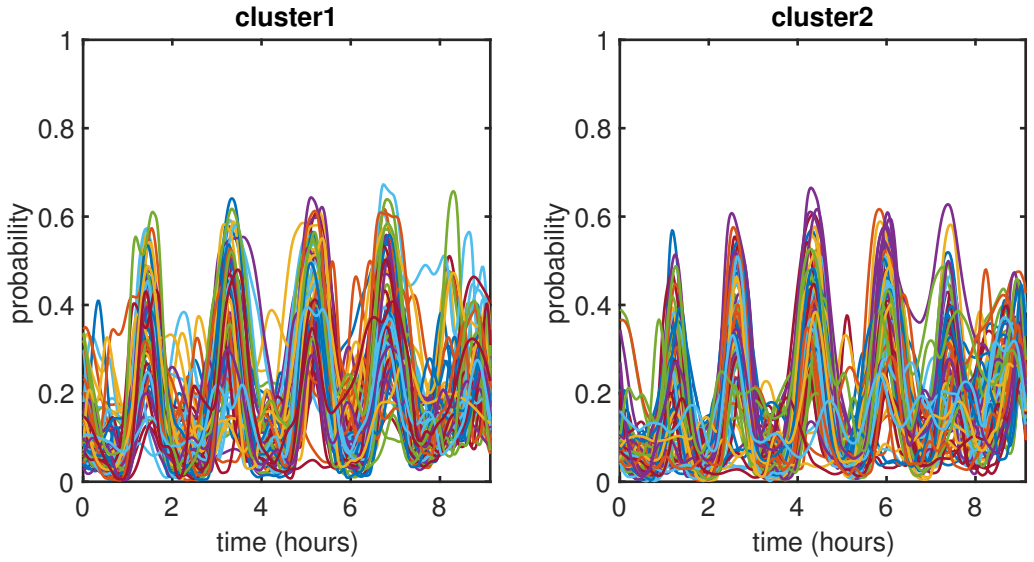


(b) The 2–step approach with the SMTW algorithm (2DTW–SMTW)

Figure 6.11: *SWS* stage. Clustering of 146 probabilistic sleep curves into two clusters by using a) the k–means clustering of misaligned curves and b) the 2–step approach with the modified SMTW algorithm in the registration step and k–medoids in the clustering step (2DTW–SMTW).



(a) The k–means clustering of misaligned sleep probabilistic curves



(b) 2–step approach with the SMTW algorithm (2DTW–SMTW)

Figure 6.12: *REM* stage. Clustering of 146 probabilistic sleep curves into two clusters by using a) the k–means clustering of misaligned curves and b) the 2–step approach with the modified SMTW algorithm in the registration step and k–medoids in the clustering step (2DTW–SMTW).

6.1.1 Relationship between sleep structure and daily measures

In this section we aim to detect whether the clusters of sleep probabilistic curves formed in the previous section significantly differ in the daily measures listed in Table 1.1. For this purpose the nonparametric testing by the Kruskal–Wallis test was selected.

	k–means	tPCS	2DTW– – SMTW	2DTW– –PCS	2DTW– –ETW
Microstate 16 (8 clusters)	FA2, age	FA2, age	FA2, age WB_m, ACT_ts	FA2, age VAS_drows	FA2, age VAS_drows
Microstate 8 (9 clusters)	pul_e, dia_m, dia_e	pul_e dia_m, dia_e ACT_sv	pul_e, dia_e, age, NMT, FA1, FA2 VAS_aff, VAS_mood VAS_drive, VAS_drows	pul_e dia_m age	pul_e, pul_m
Microstate 14 (3 clusters)	SRQ_scom, VAS_drive	SRQ_scom VAS_drive	SRQ_scom VAS_drive, VAS_drows	SRQ_scom VAS_drows, SRQ_sq FA2, dia_e	SRQ_scom VAS_drows, SRQ_sq age
Microstate 1 (2 clusters)	ACT_sv, VAS_aff	ACT_sv	ACT_sv, ACT_errp, VAS_aff, VAS_mood	ACT_sv	
Microstate 6 (3 clusters)	FA1, VAS_aff, VAS_drows	FA1, VAS_aff, VAS_drows, VAS_mood	FA1, VAS_aff VAS_mood	 VAS_aff VAS_mood	 sys_m
Microstate 13 (4 clusters)	FA3, age SRQ_sq, VAS_drows	FA3, age SRQ_sq, VAS_drows,	FA3, age SRQ_sq, VAS_drows, SRQ_scom, VAS_drive, FA2, sys_e ACT_sv, ACT_errp FMAT_l, FMAT_r	FA3, age SRQ_sq, VAS_drows,	FA3, age SRQ_sq, VAS_drows SRQ_scom, VAS_drive, FA2, pul_e
Microstate 19 (2 clusters)	age	age	age, VAS_drive	age, VAS_aff FA2	age FA2, FMAT_r, SRQ_sq, SRQ_scom

Table 6.2: Sleep microstates. List of daily measures in which significant differences between formed clusters were detected.

Wake (2 clusters)	FA2, age, SRQ_sq	FA2, age	FA2, age	FA2, age, SRQ_sq, ACT_sv, sys_e	FA2, age, SRQ_sq, ACT_sv
S1 (2 clusters)	FA1, FA2, FA3, age, SRQ_scom VAS_drive, , VAS_mood VAS_drows	FA1, FA2, FA3, age, SRQ_scom VAS_drive, VAS_mood, VAS_drows	FA1, FA2, FA3, age, SRQ_scom, VAS_drive, VAS_mood, VAS_drows	FA1, FA2, FA3, age, SRQ_scom, VAS_drive, VAS_mood, VAS_drows SRQ_sq,	FA1, FA2, FA3 age, SRQ_scom, VAS_drive, VAS_mood, VAS_drows
S2 (2 clusters)	FA2, FA3, age	FA2, FA3, age, dia_e	FA2, FA3, age SRQ_aq, ACT_errp	FA2, FA3, age	FA2, FA3, age
SWS (2 clusters)	FA2, age, SRQ_scom, ACT_sv	FA2, age, SRQ_scom, ACT_sv	FA2, age, SRQ_scom, pul_e, VAS_drows	FA2, age, SRQ_scom, ACT_ts, NMT	FA2, age
REM (2 clusters)	FA2, sys_m, WB_m, WB_e	FA2, sys_m dia_m	FA2, sys_m, ACT_ts, age, dia_e	WB_m, WB_e pul_e	FA2, sys_m, dia_m

Table 6.3: Standard sleep stages. List of daily measures in which significant differences between formed clusters were detected.

Microstates similar to *SWS*

The relationship between the structure of Microstate 16 and age or the physiological factor *FA2* was visible for both in time misaligned and aligned curves and for an arbitrary number of clusters varying between 2 and 20.

Elderly people were assigned into clusters with lower probability values of *SWS* as represented by Microstate 16, for example cluster 7 in the case of 2DTW–SMTW (Figure 6.2a) or cluster 1 when the k–means clustering was applied to the misaligned curves (Figure 6.1a). For these clusters higher values of the physiological factor *FA2* are connected with higher values of the blood pressure. On the other hand, clusters with clearly visible periods of high probabilities of Microstate 16 were formed by younger people with significantly lower values of the physiological factor.

Using the subjects' class membership obtained by 2DTW–PCS or 2DTW–ETW, the Kruskal–Wallis test detected significant differences in the level of drowsiness (*VAS_ drows*) in the morning. The subjects assigned into cluster 6 in Figure 6.3 showed significantly increased drowsiness in the morning in comparison to the subjects from cluster 4. Similarly, the subjects from cluster 2 in Figure 6.2b were more drowsy than the subjects from clusters 1, 6 or 7. Moreover, using the 2DTW–SMTW algorithm, a significant difference in the

morning well–being (WB_m) was detected between clusters 4 and 5 (Figure 6.2a).

Microstates related to *REM*

In the previous section we set the optimal number of clusters for Microstate 8 (74% *REM*) to 9. By considering the k–means clustering of misaligned curves we observed that increased probability of this microstate is associated with the increased value of the diastolic blood pressure both in the morning (dia_m) and in the evening (dia_e) and the higher pulse rate in the evening (pul_e).

After applying the 2–step algorithm, the observed difference in the evening pulse rate between formed clusters remained, but the significance of the difference in the morning or evening diastolic blood pressure varied with the choice of the selected registration method (Table 6.2). On the other hand, the 2DTW–SMTW approach detected a significant difference in the values of the physiological factor *FA2* (which represents the systolic and diastolic blood pressure both in the morning and in the evening) between clusters 8, 9 and 2 (Figure 6.4b). For the subjects assigned into cluster 2, higher values of *FA2* together with an increased probability of Microstate 8 are typical in comparison to the subjects from clusters 8 or 9.

Significant differences in the level of drive (VAS_drive), mood (VAS_mood), affectivity (VAS_aff) and drowsiness (VAS_drows) were observed. Cluster 4 in Figure 6.4b shows clearly visible oscillatory pattern of high and low probability values of Microstate 8 and the subjects from this cluster scored lower values of mood, affectivity or drowsiness in the morning indicating their better subjective feeling in contrast to the subjects from clusters 8 or 9 for which low probability values of Microstate 8 were typical. Moreover, the subjects from cluster 4 showed significantly lower values of the factor of subjectively scored sleep and awakening quality *FA1* in comparison to the subjects from cluster 8.

For the second *REM*–related Microstate 14 (72% *REM*) the optimal number of clusters was set to three. Strong relationship with the subjective assessment of the somatic complaints (SRQ_scom) and the level of drive or drowsiness in the morning was visible for both aligned and misaligned curves.

The clusters formed by the 2DTW–PCS or 2DTW–ETW method significantly differ also in the subjectively scored sleep quality questionnaire (SRQ_sq), but this effect was not observed using the k–means clustering of the misaligned curves. Similarly to Microstate 8, increased probability for Microstate 14 helps to improve sleep quality and subjective feelings in the morning. Moreover, the subjects assigned into the cluster with higher

probability values of Microstate 14 showed significantly lower somatic complaints scores (*SRQ_scom*) than the subjects with lower probability values (clusters 1 and 2 in Figure 6.5b).

Microstates similar to *S2*

Using the Kruskal–Wallis test, clusters of the sleep probabilistic curves of Microstate 1 (85% *S2*) significantly differed in the attention variability scored by the difference between extreme scores of the Alphabetical cross–out test (*ACT_sv*) [Grünberger, 1977] whether we considered misaligned curves or curves aligned by tPCS. After applying the 2DTW–SMTW algorithm, this difference remained significant but we detected also significant difference in the percentage of errors of the test (*ACT_errp*) between formed clusters. Higher variability in attention and increased percentage of errors was related with increased probability for Microstate 1.

Clusters formed by the k-means algorithm significantly differed in the level of affectivity (*VAS_aff*) in the morning. Considering the tPCS, 2DTW–PCS or 2DTW–ETW algorithm this significant difference diminished due to the changes in the cluster assignment caused by the improper alignment. On other hand, clusters formed by the 2DTW–SMTW algorithm significantly differed not only in the level of affectivity but also of mood (*VAS_mood*). We observed, that increased probability for Microstate 1 (cluster 1 in Figure 6.6b) results into impairment of mood and affectivity in the morning.

Microstates related to *Wake*

According to the results depicted in Table 6.2 we hypothesised that the structure of sleep Microstate 6 (85% *Wake*) influences the subjects' subjective feelings in the morning. Subjects with about one hour of wakefulness before the final awakening (cluster 3 in Figure 6.7a) scored higher values of mood (*VAS_mood*) or affectivity (*VAS_aff*) and they felt more drowsy in contrast to the other subjects from cluster 2 in Figure 6.7a. Similar phenomenon was observed also in the case of the factor of subjectively scored sleep quality (*FA1*). Considering the 2–step approach no new results were observed in comparison to the misaligned case. By applying 2DTW–ETW, existing relationship with *FA1* or drowsiness disappeared after the alignment (Table 6.2).

The relationship between increased probability values of Microstate 13 (45% *Wake*, 44% *S1*) and the worst subjectively scored sleep quality (*SRQ_sq*) or increased drowsiness in the morning was strong and it was visible in the case of the misaligned as well as in time

aligned curves. Moreover, the subjects above 60 years of age have higher probability values for this microstate in contrast to the younger people under 40 years of the age. On the other hand, better performance in the cognitive tests represented by the neurophysiological factor (*FA3*) was typical for clusters with lower probability values of Microstate 13. Considering the 2DTW–SMTW algorithm, similar relationship was observed also in the case of the other cognitive tests (*FMAT_l, FMAT_r, ACT_sv, ACT_errp*). Subjects with the periods of increased probability of Microstate 13 reached lower scores in the tests in comparison to the subjects with the increased probability of this microstate only at the end of the night.

The 2DTW–SMTW and 2DTW–ETW algorithms formed clusters which significantly differ in the level of drive (*VAS_drive*) in the morning as well as in the physiological factor *FA2* (Table 6.2).

Microstate 19 (88% *Wake*) is also similar to wakefulness. Considering the misaligned curves, the clusters formed by the k–means algorithm significantly differ only in age (Table 6.2). As expected, increased probability for this microstate was typical for elderly people above 60 years. After the curves alignment produced by the 2–step approach, the formed clusters differ also in daily variables representing subjective feelings in the morning like *VAS_drive* or *SRQ_sq, SRQ_scom*. Similarly to Microstates 13 or 6, the cluster with a higher probability of wakefulness scored their subjective sleep quality and somatic complaints worse than the subjects from clusters with low probability values of Microstate 19.

Standard sleep stages

In the previous section we observed that the curves alignment seems to be counterproductive for one of the wake–related sleep microstates, but for the others it helped to improve the detection of existing sleep and daily measures relationships. When considering the *Wake* stage, the k–means clustering and the 2–step approach produced similar structure of clusters (Figure 6.8). Consequently, similar significant differences between the formed clusters were detected (Table 6.3). We observed, that the cluster with higher probability values of the *Wake* stage in the first half and at the end of night (cluster 2 in Figures 6.8a and 6.8b) was formed especially by older people (above 60 years of age), while for the subjects under 40 years lower probabilities for wakefulness during the night were typical.

Moreover, the two clusters formed by an arbitrary method significantly differed in the values of the physiological factor *FA2* (Table 6.3). We observed that increased wakefulness

was related with higher values of *FA2*. Recall, that *FA2* represents mainly morning and evening blood pressure and the effect of age was removed from all daily measures (Section 1.2).

Finally, the formed clusters significantly differed in the subjectively scored sleep quality (*SRQ_sq*). As expected, higher probability for the *Wake* stage were connected with the worst sleep quality. However, after the curves alignment, either by using the 2–step approach or tPCS, this phenomenon diminished. Therefore we hypothesise that similarly to Microstate 6, the curves alignment does not bring many advantages in the case of the *Wake* stage.

The *S1* stage is very similar to the *Wake* stage and therefore we expected analogical results for these two stages. However, our expectations were not fully fulfilled. Increased probability for light sleep (cluster 1 in Figure 6.9) was related to the higher age and higher values of *FA2*. But in contrast to the *Wake* stage, slightly higher probability values for the *S1* stage during the whole night positively influence the subjective feelings in the morning (*SRQ_scom, VAS_drive, VAS_mood, VAS_drows, FA1*) and resulted into better scores in the neurocognitive tests represented by *FA3*.

In the previous section we observed that the 2DTW–SMTW approach outperformed standard clustering of misaligned curves when detecting relationships with daily measures especially when considering sleep microstates related to *S2* and *SWS*. This was true also when considering the sleep probabilistic curves mimicking the R&K stages *S2* and *SWS*.

Using any of the considered clustering method, the clusters of *S2* or *SWS* sleep probabilistic curves (Figures 6.10 and 6.11) significantly differ in *FA2* and age. In both cases slightly higher probability of the *S2* stage (cluster 1 in Figure 6.10b) or higher probability for *SWS* with typical periodic pattern and in time decreasing amplitude of peaks (cluster 1 in Figure 6.11b) were connected with lower values of *FA2* or age. This is consistent with the observations made in the case of the *Wake* or *S1* stages. In the case of the *S2* stage the formed clusters differed also in the *FA3* values which indicates that a higher probability for the stage is related to improved performance in the neurocognitive tests.

In the case of the *REM* stage, the two clusters formed by any of the considered clustering method significantly differed in values of *FA2* and also systolic blood pressure in the morning (*sys_m*). But due to the curves misalignment the clusters depicted in Figure 6.12a were difficult to interpret. Using the 2DTW–SMTW algorithm we saw, that for the first cluster a periodic pattern of the *REM* stage with four dominant in time slightly increasing peaks is typical. The subjects assigned into this cluster showed higher values

of *FA2* and *sys_ m*. Moreover, especially the subjects above 60 years of age were assigned into this cluster, the younger subjects have shown more periods of the *REM* stage (cluster 2 in Figure 6.12b). However, the difference in age was not significant when considering the clusters formed by the k–means method or tPCS.

6.1.2 Discussion and Conclusion

In this section we validated the proposed 2–step iterative approach and demonstrated that the method outperforms the truncated PCS algorithm or k–means clustering of the misaligned curves. A visually good alignment was produced especially by the 2DTW–SMTW method and this was also mirrored by the decrease of the L–criterion and AS increase.

In the 2–step approach an arbitrary method may be used for the alignment of curves within clusters, for example SMTW, PCS or ETW. All three versions of the 2–step approach worked well and led to similar results when the optimal number of clusters was set to 8 or higher. However, in the case of a smaller number of clusters, for example two or three, the 2DTW–PCS version of the 2–step algorithm produced inferior alignment. In this case, 2DTW–SMTW or 2DTW–ETW produced better results.

To achieve our main goal – to detect relationships between sleep structure and daily measures – we tested whether the differences in daily measures between formed clusters are significant or not. Relationship between *SWS* or wakefulness and age or subjectively scored sleep quality was so strong, that it was also visible when considering clusters of the original, in time misaligned, sleep curves. However and importantly, after applying the 2–step approach new correlations between sleep structure and age or daily measures were found.

First, we observed, that higher probability values for Microstate 16 (96% *SWS*, cluster 2 in Figure 6.2a) resulted in increased drowsiness in the morning. However, this phenomenon was not visible when considering misaligned sleep probabilistic curves.

When considering in time misaligned sleep probabilistic curves of the *REM*–related sleep Microstates 8 and 14, the relationship with the subjective feelings in the morning was visible only for Microstate 14. After applying the 2–step approach we observed that increased probability of Microstate 8 is related to the better subjectively scored level of mood, drive, drowsiness or affectivity in the morning.

The relationship between increased wakefulness during the night and worse subjective feelings in the morning is well known [Rosipal et al., 2013]. The cluster analysis of

misaligned sleep probabilistic curves of Microstates 6 or 13 confirmed this expected relationship. However, because of the curves misalignment, the relationship between the structure of the *Wake*–related sleep Microstate 19 and subjective feeling scores in the morning remained hidden. In contrast, after applying the 2–step approach we detected that increased probability of Microstate 19 is related to worse subjective scores of the sleep quality and increased somatic complaints.

In Rošťáková and Rosipal [2018] we observed that the curves alignment produced by the ETW algorithm is counter–productive when predicting results of daily measures by the structure of the *Wake* stage and wake–related sleep microstates. Similar results were observed for Microstate 6 representing the full wakefulness during the night and the *Wake* stage. Significant differences in *FA1*, the level of affectivity or subjectively scored sleep quality (SRQ_sq) between clusters of misaligned curves diminished after the curves alignment. It seems to be that the exact timing of the periods of increased probability of wakefulness during the night is important when the detection of existing relationships with daily measures is in the focus.

When considering the standard R&K sleep stages *Wake, S1, S2, SWS* or *REM*, we observed increased wakefulness for elderly people, while in the case of younger subjects higher probability values for the *S2* or SWS stages dominated the profiles of the clustered curves. This was observed using for both misaligned and aligned curves. The result is consistent with ageing changes in sleep described in [MedlinePlus - Health Information from the National Library of Medicine, 2018].

The statistically significant relationship between the age and *REM* stage was not observed when considering the k–means clustering of in time misaligned sleep probabilistic curves. Thanks to the curves alignment we detected that for elderly people a lower number of the *REM* stage periods is typical.

Finally, we can conclude, that the 2–step approach is a useful tool in the analysis of the sleep probabilistic curves. We hypothesise that because of its observed good performance when applied to non-trivial sleep probabilistic curves, the method can also be used in a wider range of functional data clustering tasks.

6.2 Cluster analysis of the sleep structure in patients after stroke

Investigating relationships between sleep structure of 24 patients after ischemic stroke and daily measures is in the centre of interest of this section. In contrast to the dataset of healthy sleepers the information about the patients' physiological measures; for example the pulse rate or blood pressure, was not available. In the morning, following the night with the PSG recording, the patients took part in a battery of neuropsychological tests; the Lateralised attention network test (LANT) [Greene et al., 2008], Fine motor activity test (FMAT2), Working memory test (WMT) [Kaufman and Lichtenberger, 2005] and Reaction time test (RTT) (Table 1.2). Right before and after testing they also filled in the T–MENSTAT questionnaire [Pacific Development and Technology, LCC, 2012] about subjectively scored feelings after awakening.

6.2.1 Relationship between sleep structure and daily measures

Similarly as in the case of the SIESTA database we aimed to detect existing relationships between the sleep structure and the results of daily measures of patients after ischemic stroke. In this case we used the 2–step approach with the modified SMTW algorithm in the registration step (2DTW–SMTW).

In the case of healthy sleepers we observed that for one wake–related sleep microstate and for the *Wake* stage the curves alignment led to inferior results. Therefore the k–means clustering of misaligned sleep probabilistic curves was considered as well.

Because of a small number of subjects we varied the number of clusters between two and 10. The optimal number of clusters was again selected by using the AS and L–criterion.

However, not all subjects took part in the whole battery of daily tests, for example the complete results of the LANT were available only for 19 subjects or only 17 subjects performed the complete FMAT2. Therefore the procedure for finding the optimal number of clusters and the cluster analysis were repeated for each daily measure separately by considering only the patients with non–missing values.

	k–means	2DTW–SMTW
Microstate 1 (2 clusters)	LANT_RVF_OF	FMAT2_cp1, FMAT2_cp2 FMAT2_cp4
Microstate 16 (3 clusters)		FMAT2_cp1, FMAT2_sp1
Microstate 6 (3 clusters)		FMAT2_sp3, FMAT2_sp6
Microstate 5 (4 clusters)		LANT_A
Microstate 10 (2 clusters)	RTT_3	RTT_3
Microstate 4 (2 clusters)	LANT_OI, WMT_bw T–MENSTAT_A_1	LANT_OI, WMT_bw, T–MENSTAT_A_1
Microstate 12 (2 clusters)	RTT_3	RTT_3
Microstate 11 (2 clusters)	RTT_min, RTT_4, LANT_C, LANT_LVF_A, LANT_LVF_OF T–MENSTAT_B_3,4	RTT_min, RTT_4
Microstate 19 (2 clusters)	NIHSS	NIHSS
S1 (2 clusters)	RTT_min, RTT_4 FMAT2_sp4, WMT_bw	
S2 (2 clusters)	LANT_A, LANT_RVF_A	LANT_A, LANT_RVF_A
SWS (2 clusters)	WMT_bw	WMT_bw T–MENSTAT_B_2,3,4

Table 6.4: List of daily measures in which significant differences between formed clusters of patients after stroke were detected.

Sleep microstates similar to the *Wake* stage or light sleep

Sleep Microstate 1 (50% *S1*) represents light sleep. According to the values of the AS and L–criterion, the optimal number of clusters was set to two (Figure 6.15a) for all daily measures. Only 17 patients after stroke performed the whole battery of the Fine motor activity test (FMAT2). The 2DTW–SMTW approach formed two clusters which significantly differ in the number of correctly retraced pixels in patterns 1, 2 and 4 (*FMAT2_ cp1, FMAT2_ cp2, FMAT2_ cp4*) (Figure 1.3). The three subjects from cluster 1 in Figure 6.15a with higher probability values for the microstate showed significantly more correctly redrawn points in comparison to 14 subjects from the second cluster.

Similar result was observed for Microstate 16 (51% *S1*). The optimal number of clusters was set to three. Ten subjects from the first cluster in Figure 6.15b with close to zero probability values during the whole night were not able to retrace the pattern 1 as successfully as their six colleagues from cluster 2. The third cluster includes one subject with possibly an outlier profile.

Three clusters were set as optimal also for Microstate 6 (43% *S1*, 44% *S2*). However, in comparison to Microstates 1 and 16, different results for the FMAT2 were observed. Four subjects assigned into the second cluster (Figure 6.15c) showed significantly fewer successively redrawn pixels in patterns 3 and 6 in contrast to six subjects from the first cluster.

For Microstate 5 (68% *S1*) and 19 subjects who completed the LANT the optimal number of clusters was set to four. The subjects assigned into the first and second cluster in Figure 6.13 reached negative (and therefore better) LANT_A scores and significantly differed from the subjects from clusters 3 and 4 with positive scores in the test.

Finally we observed a relationship between the increased probability for Microstate 10 (59% *S1*) and increased reaction time (Table 6.4).

Sleep microstates similar to the *S2* or *SWS* stage

The sleep probabilistic curves for Microstate 4 (28% *S1*, 42% *S2*, 28% *SWS*) of 23 patients who took part in the Working memory test (WMT) were divided by 2DTW–SMTW into two clusters (Figure 6.14). We observed, that increased probability for the microstate led to better scores in the backward part of the WMT (*WMT_ bw*) as well as in LANT_OI. On the other hand, higher probability values for the microstate were connected with subjectively decreased level of energy or motivation before the neurocognitive testing as reflected by lower scores in T–MENSTAT_A_1.

The structure of sleep Microstates 12 (44% *S2*, 31% *SWS*) and 11 (65% *SWS*) influence subject's reaction times, but in different ways. In both cases the optimal number of clusters was set to two for 23 subjects. The patients after stroke with increased probability for Microstate 12 showed shorter reaction times, on the other hand the higher probability for Microstate 11 resulted into increased reaction times (Table 6.4).

After performing the whole battery of neurocognitive tests only 19 patients filled in the T–MENSTAT questionnaire (T–MENSTAT_B). When considering Microstate 11, the structure of clusters formed by the k–means algorithm and 2DTW–SMTW differed only in the assignment of one curve (Figure 6.17). It is difficult to visually decide, whether the green curve in Figure 6.17a belongs to the second cluster or its profile is similar to curves from cluster 1. However, the second cluster produced by the k–means clustering showed significantly higher subjective level of frustration and drowsiness (T–MENSTAT_B_3, T–MENSTAT_B_4). This effect disappeared after the 2DTW–SMTW clustering. Therefore we speculate that the observed relationship between the structure of Microstate 11 and the level of frustration and sleepiness was a random. To confirm this speculation a wider cohort of patients after stroke would be needed.

Considering the k–means clustering of the sleep probabilistic curves for Microstate 11, the clusters significantly differed in several components of the LANT (Table 6.4). However, the p–values approximately equal to 0.046 indicate that the significance of the differences may be caused only by a chance and the assignment of several misaligned curves into an incorrect cluster. This observation was confirmed by the 2DTW–SMTW approach, where no significant differences in the LANT components where observed between clusters.

Finally, Microstate 19 (86% *SWS*) reflects the severity of stroke. The optimal number of clusters was chosen as two. The subjects with higher NIHSS values and therefore with more severe stroke were assigned into the cluster with higher probability values for this microstate (Figure 6.18).

Standard sleep stages

In addition to sleep microstates we performed the cluster analysis of sleep probabilistic curves mimicking the standard R&K sleep stages. Recall, that in the case of patients after stroke only probabilistic curves for the *Wake* stage and the *nonREM* stages were available (Section 2.3.2).

There was no difference in daily measures observed between clusters based on the structure of the *Wake* stage.

In the case of the *S1* stage the optimal number of clusters was set to two. The Kruskal–Wallis test found no significant differences in daily measures between clusters formed by 2DTW–SMTW. On the other hand, considering the k–means clustering of misaligned curves, the subjects assigned into the cluster with higher probability values for light sleep showed significantly higher reaction times (*RTT_min, RTT_4*) and worst scores in the WMT (*WMT_bw*). On the other hand, they performed better in the FMAT2 (*FMAT2_sp4*) in comparison to their colleagues with lower probability values for the *S1* stage (Figure 6.19).

Considering the character of the *S2* stage, the two clusters depicted in Figure 6.20 significantly differ in the results of LANT_A and LANT_RFV_A. The second cluster with higher probability values for the *S2* stage across the whole night includes subjects with lower and therefore better scores in both LANT_A and LANT_RVF_A components of the LANT.

Considering *SWS*, two clusters formed by 2DTW–SMTW (Figure 6.21) significantly differ in the results of the backward part of the WMT (*WMT_bw*) and the T–MENSTAT questionnaire filled in after the neurocognitive testing (T–MENSTAT_B_2,3,4). Subjects assigned into the cluster with higher probability values for *SWS* remembered, on average, more digits in the backward order and felt less drowsy, exhausted or frustrated after testing than their colleagues with lower probability values for *SWS*.

6.2.2 Discussion and Conclusion

In this section we applied the 2–step approach for curves clustering and alignment on the dataset of patients after stroke. We observed that increased probability of sleep microstates similar to light sleep (Microstates 1 and 16) helped to improve performance in the FMAT2. Similar result was observed when considering the curves representing the *S1* stage. On the other hand, increased probability for Microstate 6, representing the "border" between stages *S1* and *S2*, leads to the opposite results.

Considering light sleep we also observed that subjects with higher probability values for sleep Microstate 5 performed better in the LANT_A component of the LANT. When considering the standard R&K sleep stages the only relationship with the LANT_A component was observed for the *S2* stage.

Increased probability for microstates laying on the border between the *S2* stage and *SWS* lead to the improved reaction times while the structure of microstates related only to *SWS* influences the reaction speed in a negative way. Moreover, sleep Microstate 4 or

SWS showed relationships with the results of the backward part of the WMT and the T–MENSTAT questionnaire. The subjects with increased *SWS* during the whole night felt subjectively less drowsy, exhausted or irritated after performing the whole battery of neurocognitive tests.

Finally we observed that higher probability values for Microstate 19 similar to *SWS* are typical for subjects following more severe stroke (higher NIHSS).

The benefit of the 2–step approach in comparison to the k–means clustering of in time misaligned curves was not evident. This may be due to a small number of subjects and clusters. Considering the *S2* or *SWS* stages and with them associated sleep microstates, the k–means clustering of misaligned curves produced the same results as the 2DTW–SMTW approach.

In the case of the *S1* stage we observed that similarly to the clustering of the *Wake* stage curves of healthy sleepers, the time alignment seems to be counterproductive. In other words, the exact occurrence of the periods of light sleep is important when detecting relationships with the studied daily features. However, we observed the opposite phenomenon in the case of sleep microstates related to light sleep, where the time alignment helped to detect otherwise hidden relationships. Note that these analysed microstates show not negligible weights also for the *S2* stage (Microstates 6, 10 or 15).

Overall, we need to stress that due to a small number of patients these observations and conclusions can be viewed only as preliminary.

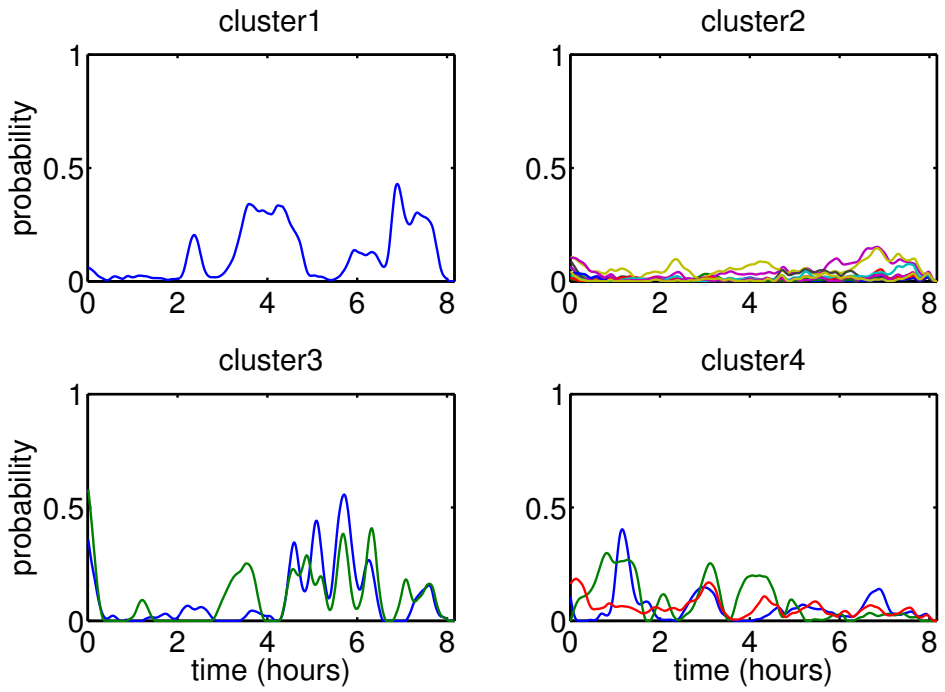


Figure 6.13: Microstate 5. Clustering of sleep probabilistic curves into four clusters by the 2DTW–SMTW algorithm. 19 patients who completed the LANT were included in the analysis.

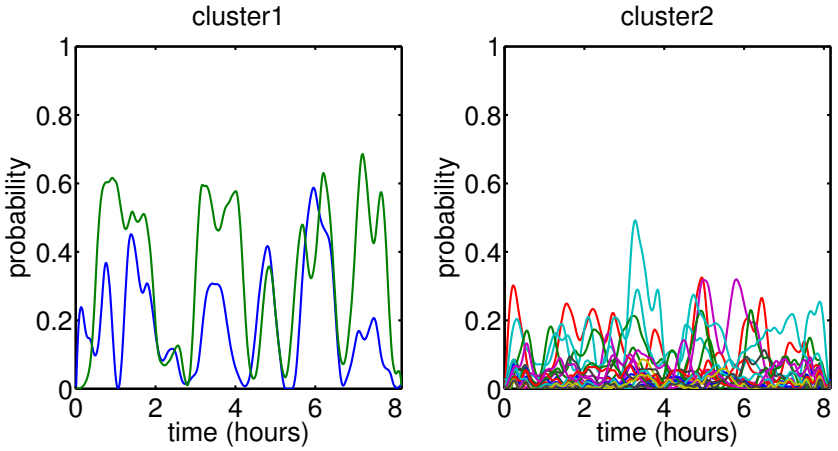
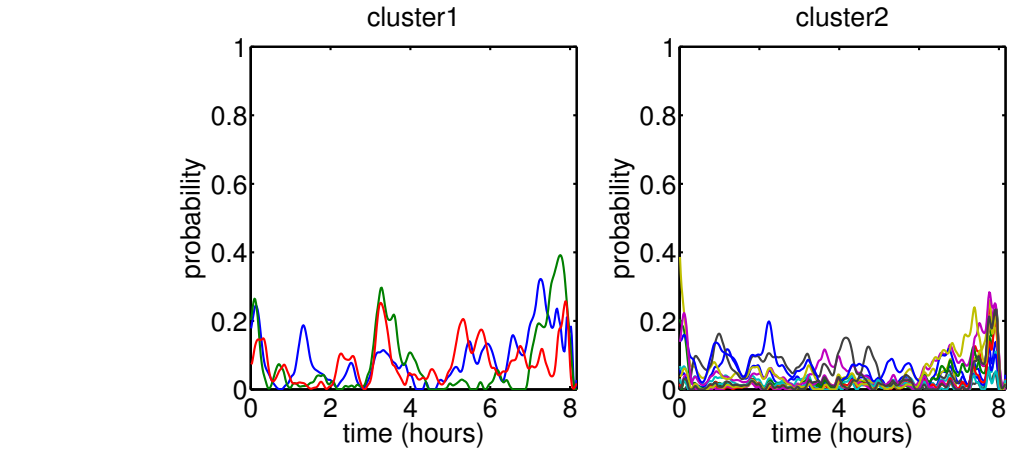
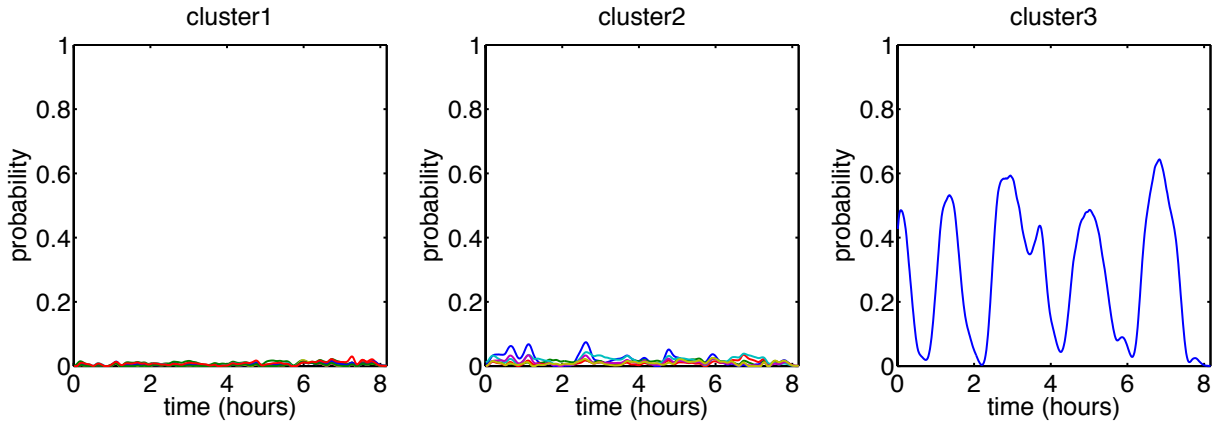


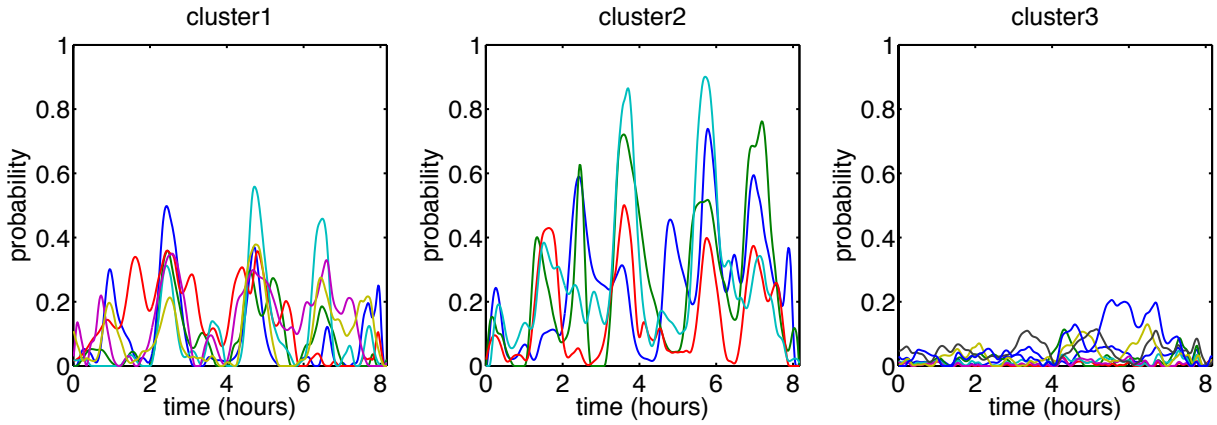
Figure 6.14: Microstate 4. Clustering of sleep probabilistic curves into two clusters by the 2DTW–SMTW algorithm. 23 patients who completed the Working memory test were included in the analysis.



(a) Microstate 1

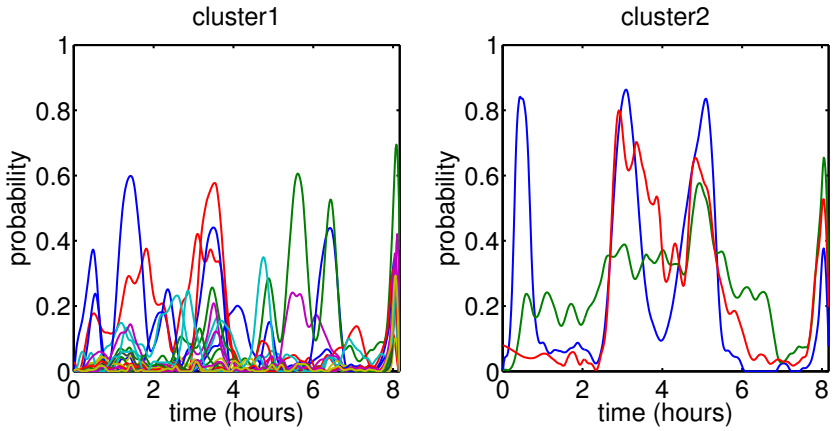


(b) Microstate 16

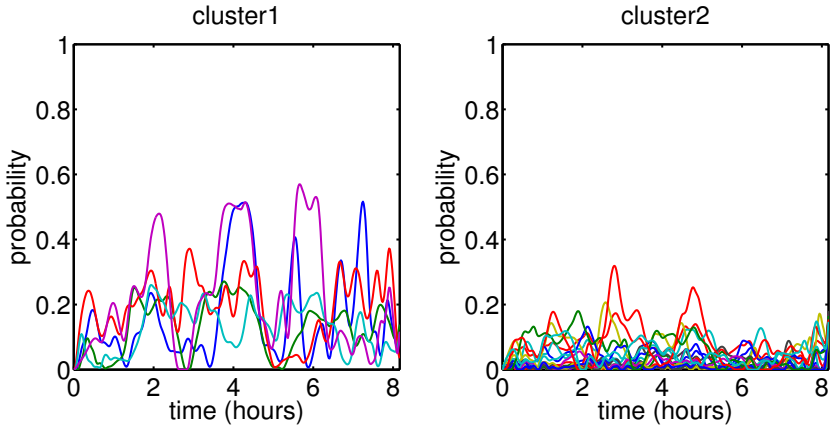


(c) Microstate 6

Figure 6.15: Microstates 1, 16 and 6. Clustering of sleep probabilistic curves representing light sleep by the 2DTW–SMTW algorithm. 17 patients who completed the Fine motor activity test were included in the analysis. The optimal number of clusters for Microstate 1 was set to two, for Microstates 16 and 6 it was set to three.

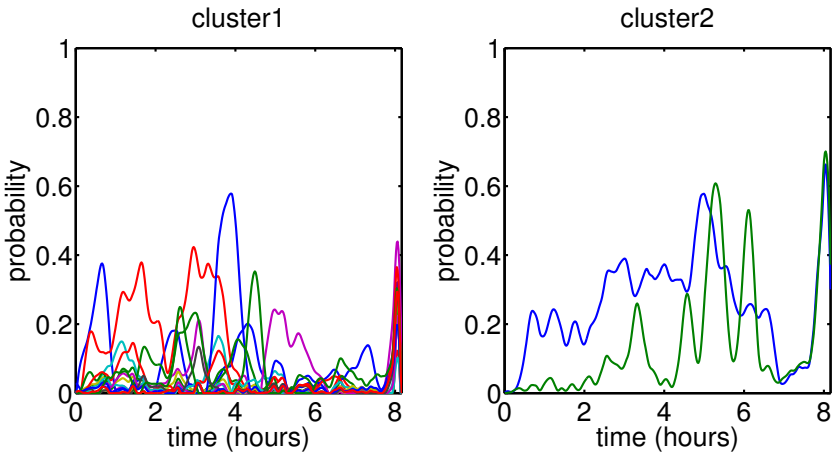


(a) Microstate 11

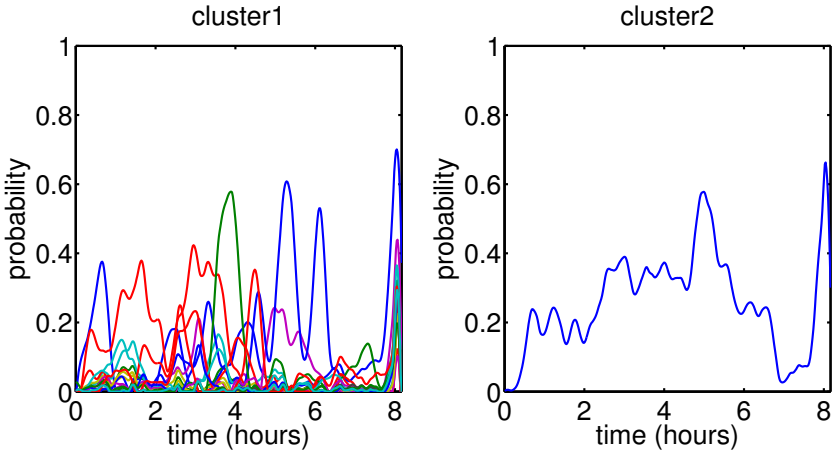


(b) Microstate 12

Figure 6.16: Microstates 11 and 12. Clustering of sleep probabilistic curves into two clusters by the 2DTW–SMTW algorithm. 23 patients who completed the Reaction time test were included in the analysis.



(a) k–means



(b) 2DTW–SMTW

Figure 6.17: Microstate 11. Clustering of sleep probabilistic curves into two clusters by the k–means algorithm or 2DTW–SMTW. 19 patients who filled in the T–MENSTAT_B questionnaire were included in the analysis.

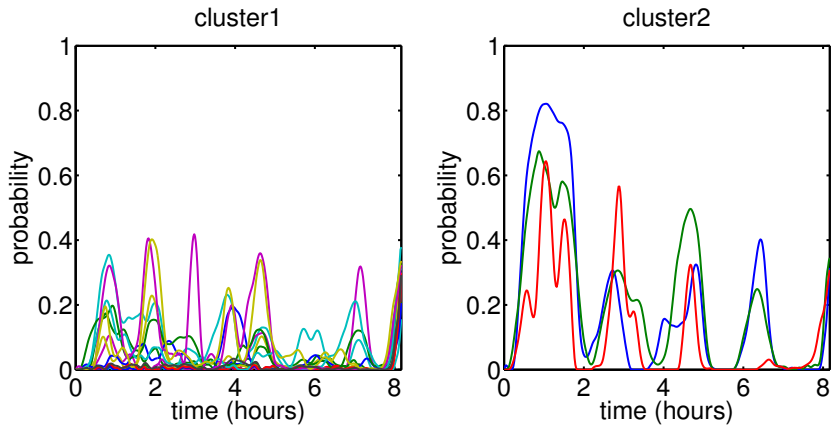


Figure 6.18: Microstate 19. Clustering of 24 probabilistic sleep curves into two clusters by the 2DTW–SMTW algorithm.

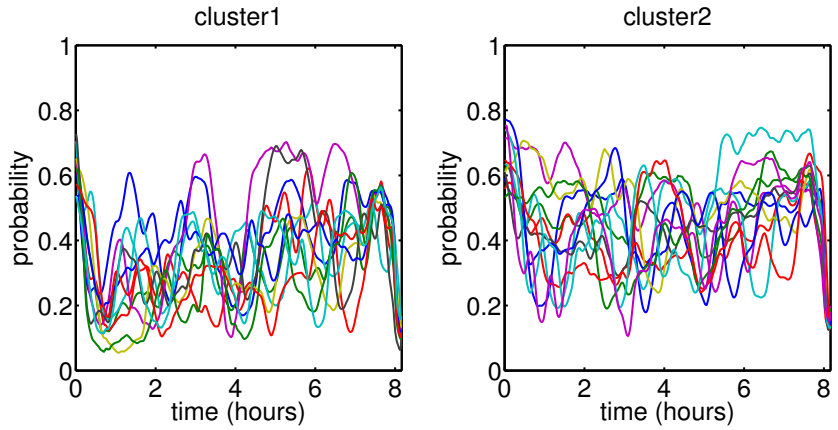


Figure 6.19: *S1* stage. The k–means clustering of in time misaligned sleep probabilistic curves into two clusters. 23 patients who completed the Working memory test were included in the analysis.

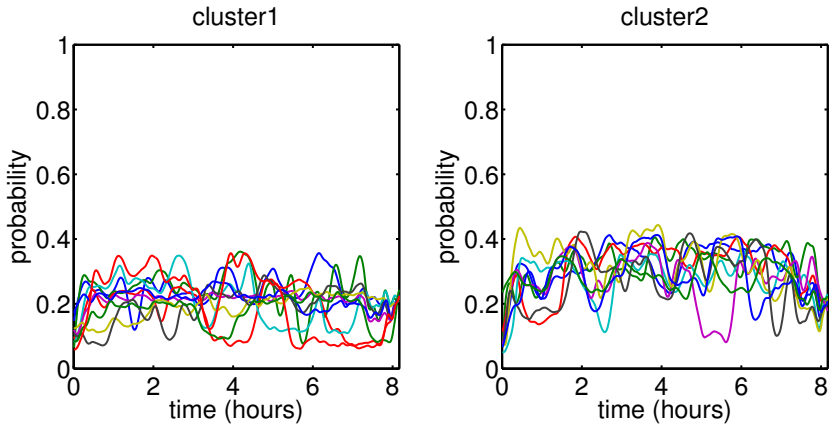


Figure 6.20: *S2* stage. Clustering of sleep probabilistic curves into two clusters by the 2DTW–SMTW algorithm. 19 patients who completed the LANT were included in the analysis.

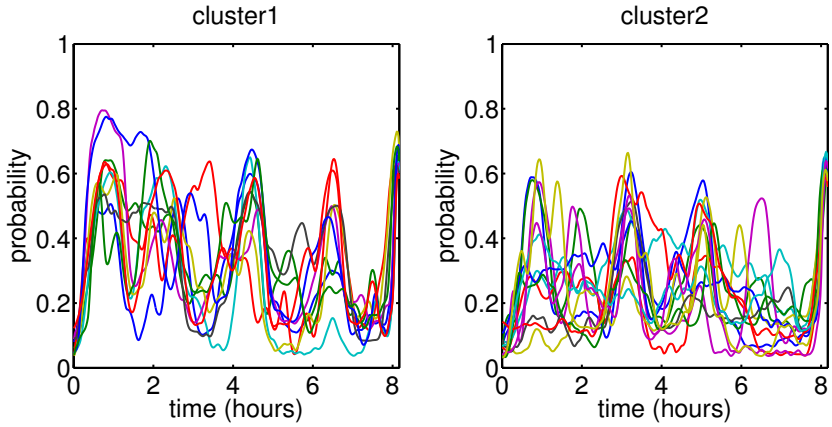


Figure 6.21: *SWS*. Clustering of sleep probabilistic curves into two clusters by the 2DTW–SMTW algorithm. 23 patients who completed the Working memory test were included in the analysis.

Chapter 7

Multilevel functional principal component analysis

Multilevel functional data are data observed over a continuum, for example time, with repeated measurements. In the SIESTA database [Klösch et al., 2001] for each subject the PSG recordings of the two consecutive nights are available. Similarly, the results of cluster analysis of the sleep probabilistic curves can be viewed as an example of multilevel functional data. Now, each cluster is an "object" and curves assigned into this cluster are repeated observations of the object.

Functional principal component analysis (FPCA) is the key technique for dimensionality reduction and detection of main directions of variability present in functional data. However, it is not the most suitable tool for the situation when analysed dataset contains repeated or multiple observations, because information about repeatability of measurements is not taken into account. Multilevel functional principal component analysis (MFPCA) [Di et al., 2009] is the modified version of FPCA developed for data observed at multiple visits or measurements.

The chapter is organised as follows: First a short introduction of FPCA is provided. More attention is paid to MFPCA, its modifications and application to sleep data.

7.1 Functional principal components analysis

FPCA is the method developed for finding the strongest directions of variability included in functional data and for the dimensionality reduction. Let consider a square–integrable random function X defined on the time interval $T = [0, 1]$ with mean $\mu : T \to \mathbb{R}$ and

covariance function $R : T \times T \to \mathbb{R}$. Using the Mercer's theorem [Mercer, 1909] the spectral decomposition of R follows the formula

$$R(s,t) = \sum_{i=1}^{\infty} \lambda_i \phi_i(t) \phi_i(s),$$

where $\lambda_1, \lambda_2, \dots$ are eigenvalues and $\phi_1, \phi_2, \dots$ are eigenfunctions of R satisfying

$$\int_T R(s,t)\phi_i(t)dt = \lambda_i \phi_i(s), \quad i = 1, 2, \dots,$$

$$\int_T \phi_i(t)\phi_j(t)dt = \begin{cases} 0, & i \neq j, \\ 1, & i = j. \end{cases}$$

While $\{\phi_k\}_{k=1}^{\infty}$ forms functional basis of the $\mathrm{L}^2[0,1]$ space, the Karhunen–Loewe expansion [Karhunen, 1947; Loeve, 1945] of a random function X is the following

$$X(t) = \mu(t) + \sum_{k=1}^{\infty} \alpha_k \phi_k(t), t \in T. \tag{7.1}$$

Then $\{\phi_k\}_{k=1}^{\infty}$ are functional principal components and the principal component scores coefficients $\alpha_k, k = 1, 2, \dots$ are considered as independent random variables with zero mean and variance λ_k. It is easy to show that

$$\alpha_k = \int_T (X(t) - \mu(t))\, \phi_k(t)dt, \;\; k = 1, 2, \dots. \tag{7.2}$$

In practice, the estimators for μ and R are usually found by the method of moments applied to the data. Then the functional principal components are estimated as the eigenfunctions of the covariance function estimator $\widehat{R}$. Finally, the principal component scores are estimated by the formula (7.2) by replacing μ and $\phi_1, \phi_2, \dots$ by their estimations.

In the Principal component analysis (PCA) the number of principal components is equal or less than the dimension of the vector space we are working in. In the case of the functional data we are in the space with possibly infinite dimension. On the other hand, if our database included N independent curves, then at most $N - 1$ eigenfunctions of R are nonzero [Ramsay and Silverman, 2005].

In fact, only the several functional principal components explain the majority of the data variability. An optimal number P of the functional principal components can be chosen similarly as in the case of PCA, for example such that the variance explained by the first P functional principal components is above a given threshold κ

$$\frac{\sum_{i=1}^{P} \lambda_i}{\sum_{j=1}^{N-1} \lambda_j} > \kappa.$$

In this work we set $\kappa = 0.9$. More approaches for the selection of P can be found in [Ramsay and Silverman, 2005].

7.2 Multilevel functional principal component analysis in details

Multilevel FPCA (MFPCA) [Di et al., 2009] is a version of FPCA developed for functional data with repeated measurements. In other words, for each subject at least two observations are available. This method distinguishes two types of variability – the variability between and within subjects. As an example we can mention the sleep probabilistic curves of a given sleep microstate from the first and the second night.

Let consider I subjects with repeated observations defined on a closed time interval T. As in the previous section, we assume for simplicity $T = [0, 1]$. To follow the ideas from [Di et al., 2009] we will assume the balanced design – it means that for each subject the same number of observations $J > 1$ is available and the observations within subjects have a natural order; for example nights in our database of healthy sleepers.

Let us denote the j^{th} observation of the i^{th} subject in time $t \in T$ by $X_{ij}(t)$. A two–way functional ANOVA model with random effects is used to model the structure of the data

$$X_{ij}(t) = \mu(t) + \eta_j(t) + Z_i(t) + W_{ij}(t), \;\; i = 1, \ldots, I, \;\; j = 1, \ldots, J. \tag{7.3}$$

Here, μ is the overall mean function and $\eta_j, j = 1, \ldots, J$ is the observation–specific deviation from the overall mean satisfying $\sum_{j=1}^{J} \eta_j(t) = 0, \;\; \forall \;\; t \in T$ for identifiability. These functions are considered as fixed effects and their estimators

$$\widehat{\mu}(t) = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} X_{ij}(t) = \overline{X}_{\cdot\cdot}(t),$$

$$\widehat{\eta_j}(t) = \frac{1}{I} \sum_{i=1}^{I} X_{ij}(t) - \overline{X}_{\cdot\cdot}(t) = \overline{X}_{\cdot j}(t) - \overline{X}_{\cdot\cdot}(t)$$

are similar to their versions from the standard ANOVA model.

Random effects $Z_i, i = 1, \ldots I$ represent the subject–specific deviations from the observation–specific mean and W_{ij} is the residual deviation from the subject– and observation–specific profile. Z_i and W_{ij} are considered to be zero–mean stochastic processes defined over a common probability space $(\Omega, \mathcal{S}, \mathcal{P})$ with adequately smooth covariance functions R_Z :

$T \times T \to \mathbb{R}$ and $R_W : T \times T \to \mathbb{R}$. Moreover, Z_i and W_{ij} are uncorrelated for each $i = 1, \ldots, I$ and $j = 1, \ldots, J$.

Using the Karhunen–Loewe expansion the random effects can be rewritten into the form

$$Z_i(t) = \sum_{k=1}^{\infty} \alpha_{ik} \phi_k^{(1)}(t), \quad i = 1, \ldots, I, \tag{7.4}$$

$$W_{ij}(t) = \sum_{k=1}^{\infty} \beta_{ijk} \phi_k^{(2)}(t), \quad i = 1, \ldots, I; \quad j = 1, \ldots, J. \tag{7.5}$$

Here, $\{\phi_k^{(1)}\}_{k=1}^{\infty}$ and $\{\phi_k^{(2)}\}_{k=1}^{\infty}$ are the eigenfunctions of R_Z and R_W respectively and they are called the level 1 and level 2 eigenfunctions or functional principal components. Each set of eigenfunctions forms an orthogonal basis of the $L^2[0,1]$ functional space, but the two functional bases are not necessarily mutually orthogonal. Coefficients $\{\alpha_{ik}\}_{k=1}^{\infty}$ and $\{\beta_{ijk}\}_{k=1}^{\infty}$ are random variables with zero mean and

$$\mathrm{E}(\alpha_{ik}\alpha_{il}) = \begin{cases} 0, & \text{if } k \neq l, \\ \lambda_k^{(1)}, & \text{if } k = l. \end{cases} \qquad \mathrm{E}(\beta_{ijk}\beta_{ijl}) = \begin{cases} 0, & \text{if } k \neq l, \\ \lambda_k^{(2)}, & \text{if } k = l. \end{cases}$$

We call them the level 1 and level 2 principal component scores. Moreover, $\{\alpha_{ik}, k = 1, 2, \ldots\}$ are assumed to be uncorrelated with $\{\beta_{ijl}, l = 1, 2, \ldots\}$ to mirror uncorrelation between Z_i and W_{ij}.

In order to estimate eigenfunctions and principal component scores on both levels we need to find an appropriate estimators for R_Z and R_W. Using the model (7.3) and the formulas (7.4) and (7.5) it is not difficult to show that

$$\begin{aligned} &\operatorname{cov}(X_{ij}(s), X_{ij}(t)) = R_Z(s,t) + R_W(s,t), \\ &\operatorname{cov}(X_{ij}(s), X_{ik}(t)) = R_Z(s,t), \qquad i = 1, \ldots, I; \quad j, k = 1, \ldots, J; j \neq k; \quad s, t \in T. \end{aligned}$$

Let $R_T = R_Z + R_W$ be the total covariance function. Di et al. [2009] estimated the unknown covariance functions by the method of moments

$$\widehat{R}_T(s,t) = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \left(X_{ij}(s) - \widehat{\mu}(s) - \widehat{\eta}_j(s)\right)\left(X_{ij}(t) - \widehat{\mu}(t) - \widehat{\eta}_j(t)\right), \tag{7.6}$$

$$\widehat{R}_Z(s,t) = \frac{1}{IJ(J-1)} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{l \neq j}^{J} \left(X_{ij}(s) - \widehat{\mu}(s) - \widehat{\eta}_j(s)\right)\left(X_{il}(t) - \widehat{\mu}(t) - \widehat{\eta}_l(t)\right), \tag{7.7}$$

$$\widehat{R}_W(s,t) = \widehat{R}_T(s,t) - \widehat{R}_Z(s,t). \tag{7.8}$$

The proposed estimators are asymptotically unbiased estimators of the covariance func-

tions R_Z and R_W for $I \to \infty$

$$\begin{aligned}
\mathrm{E}\left(\widehat{R}_T(s,t)\right) &= \frac{I-1}{I} R_T(s,t) \xrightarrow{I\to\infty} R_T(s,t), \\
\mathrm{E}\left(\widehat{R}_Z(s,t)\right) &= \frac{I-1}{I} R_Z(s,t) \xrightarrow{I\to\infty} R_Z(s,t), \\
\mathrm{E}\left(\widehat{R}_W(s,t)\right) &= \frac{I-1}{I} R_W(s,t) \xrightarrow{I\to\infty} R_W(s,t).
\end{aligned}$$

If we want to be really strict, then it is possible to multiply each estimator by $\frac{I}{I-1}$ to obtain unbiased estimators. However, this would cause no changes in the estimated level 1 or level 2 functional principal components and only estimated eigenvalues would be multiplied by a constant.

Then the level 1 and the level 2 functional principal components are obtained as the eigenfunctions of $\widehat{R}_Z$ and $\widehat{R}_W$. Similarly to PCA or FPCA only finite number of the level 1 and the level 2 eigenvalues is nonzero. Di et al. [2009] chooses the optimal numbers P_1, P_2 of the functional principal components on level 1 or 2 by the following criterion

$$P_i = \min\left\{ k : \frac{\sum_{j=1}^{k} \lambda_j^{(i)}}{\sum_{j=1}^{IJ-1} \lambda_j^{(i)}} \geq V_1 \ \wedge \ \lambda_k^{(i)} < V_2 \right\}, \quad i = 1, 2,$$

where V_1, V_2 are the cumulative variance explained thresholds. Di et al. [2009] set $V_1 = 0.9$ and $V_2 = \frac{1}{n}$, where n is equal to the number of time points where the curves are observed.

In FPCA, the principal component scores can be derived directly from the formula (7.2). However, this is not possible in MFPCA due to the fact that the level 1 and level 2 functional principal components are not necessarily mutually orthogonal

$$\begin{aligned}
\int_T \left(X_{ij}(t) - \mu(t) - \eta_j(t)\right) \phi_k^{(1)}(t) dt &= \alpha_{ik} + \sum_{l=1}^{\infty} \beta_{ijk} \int_T \phi_k^{(1)}(t) \phi_l^{(2)}(t) dt, \\
\int_T \left(X_{ij}(t) - \mu(t) - \eta_j(t)\right) \phi_l^{(2)}(t) dt &= \beta_{ijl} + \sum_{k=1}^{\infty} \alpha_{ik} \int_T \phi_k^{(1)}(t) \phi_l^{(2)}(t) dt.
\end{aligned}$$

Di et al. [2009] use a linear mixed effects model for estimation of the principal component scores. The exact methodology is described in details in [Di et al., 2009] and therefore it is omitted here.

7.2.1 Application to the sleep dataset

In this section the MFPCA algorithm is applied to the sleep data from the SIESTA database (Section 1.2). In the SIESTA database for each subject information about the

first and the second night spent in the sleep lab are available. Therefore the data are appropriate for the analysis provided by the standard MFPCA method.

The MFPCA algorithm was applied to the sleep probabilistic curves of a given sleep microstate or sleep stage mimicking the R&K stages. We considered either original sleep probabilistic curves or the sleep probabilistic curves aligned within each subject separately. Because of a small number of observations per subject and a high level of similarity between the sleep probabilistic curves within a subject, the modified SMTW method worked well.

The MFPCA model has the form

$$X_{ij}(t) = \mu(t) + \eta_j(t) + Z_i(t) + W_{ij}(t), \qquad i = 1, \ldots, 146, \;\; j = 1, \ldots, 2.$$

First, we analysed the night effect of each sleep microstate or sleep stage separately. In the second step we computed the coefficient of multiple correlation between the difference in two nights of a subject and the corresponding difference in daily measures specified in Table 1.1.

The effect of night

First, the night–specific profiles $\delta_j(t) = \mu(t) + \eta_j(t), j = 1, 2$ were estimated for each sleep microstate or sleep stage separately. Because the results for in time aligned and misaligned sleep probabilistic curves were approximately the same, we analysed only the first one.

The difference between the first and second night–specific profiles is negligible within the majority of the sleep microstates (Figure 7.1). Considering Microstates 13 (45% *Wake*, 41% *S1*) and 19 (88% *Wake*), for the first night a slightly higher probability values are typical. On the contrary, higher probability values of Microstate 14 (72% *REM*) are typical for the second night.

More visible differences between the night–specific profiles were detected for the sleep stages. When a subject sleeps for the first time in a new environment, his or her sleep is lighter and problems with falling asleep occur. This phenomenon is called the "first–night effect" [Roth et al., 2005; Tamaki et al., 2016] and it is visible in Figure 7.2. The probability values for the *Wake* stage are higher for the first night in comparison to the second night. On the contrary, for the second night the sleep probabilistic curves of stages *S2* and *REM* lie higher. The difference between the nights profiles in the *SWS* stage is visible especially at the beginning of the night, for the second night the first two periods of *SWS* reach higher local maxima values.

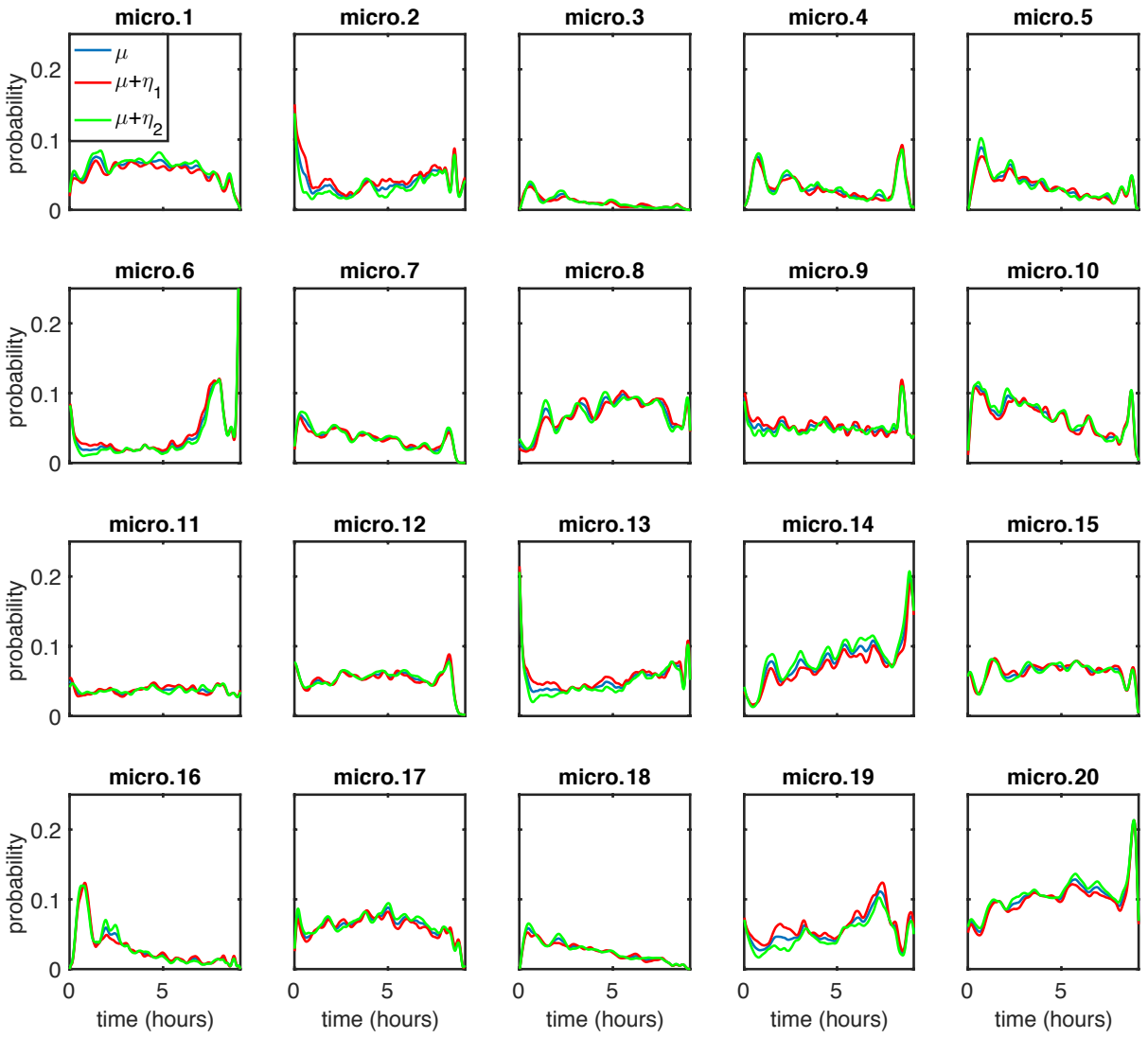


Figure 7.1: Night profiles for 20 sleep microstates estimated by the MFPCA method. Before applying the MFPCA method the curves were aligned with the modified Self–modelling time warping method (SMTW) within each subject separately. The overall mean function μ for each microstate is depicted in blue, red curve represents the first night effect $\mu + \eta_1$ and green curve represents the second night effect $\mu + \eta_2$.

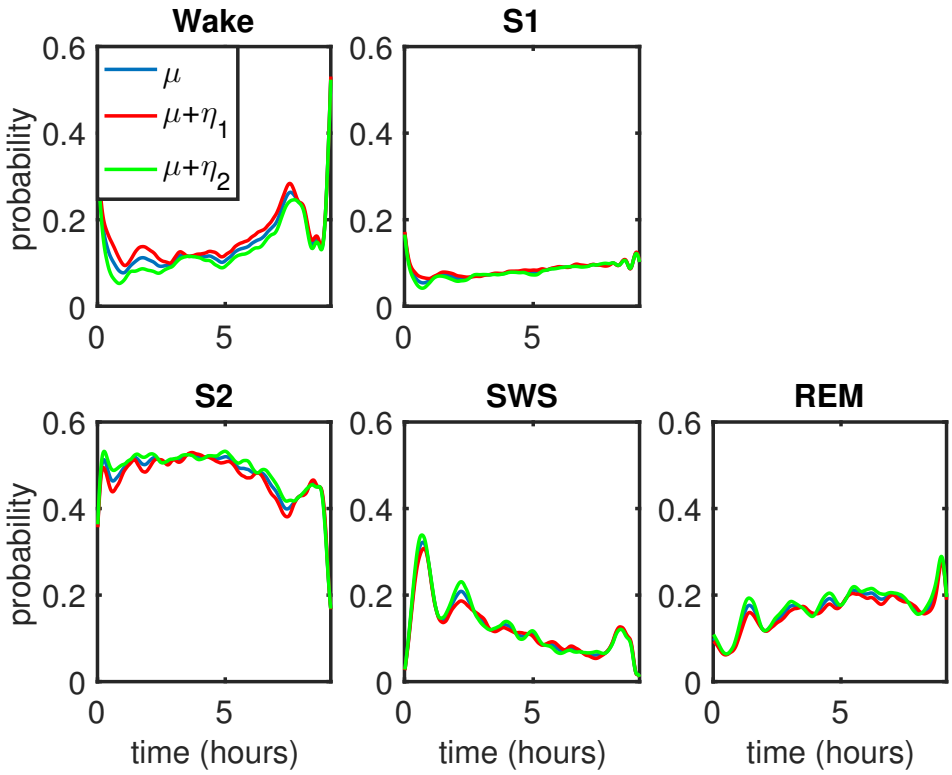


Figure 7.2: Night profiles of sleep stages mimicking the standard R&K staging and estimated by the MFPCA method. Before applying the MFPCA method the curves were aligned with the modified Self–modelling time warping method (SMTW) for each subject separately. The notation is the same as in Figure 7.1.

Correlations between differences in the night profiles and daily measures

In the second step, we investigated whether a difference between nights of a subject is mirrored also in a difference in a daily measure. For this purpose we used the coefficient of multiple correlation.

Let d to be a zero–mean random variable representing the difference in a daily measure. The random vector $\beta_{diff} = (\beta_{11} - \beta_{21}, \ldots, \beta_{1P_2} - \beta_{2P_2})^T$ represents the difference in the level 2 principal component scores and according to [Di et al., 2009]

$$\mathrm{E}(\beta_{diff}) = 0 \in \mathbb{R}^{P_2} \qquad \text{and} \qquad \mathrm{Cov}(\beta_{diff}) = 2\Lambda^{(2)},$$

where $\Lambda^{(2)}$ is a diagonal matrix with the first P_2 level 2 eigenvalues on its main diagonal. Consequently the correlation matrix of β_{diff} is equal to the identity matrix $\mathbb{I}_{P_2}$. Finally,

let γ to be a P_2–dimensional vector of correlation coefficients between the variable d and each component of the random vector β_{diff}, that is $\gamma_i = \rho\left(d, \beta_{diff_i}\right)$.

The coefficient of multiple correlation, which we denote ρ_m, measures how well a single variable d can be predicted using a linear combination of a set of variables β_{diff}. In other words

$$\rho_m\left(d, \beta_{diff}\right) = \max_{a \in \mathbb{R}^{P_2}} \rho\left(d, a^T \beta_{diff}\right),$$

which can be simplified into the formula

$$\rho_m^2\left(d, \beta_{diff}\right) = \gamma^T \left(\mathbb{I}_{P_2}\right)^{-1} \gamma = \gamma^T \gamma = \sum_{k=1}^{P_2} \gamma_k^2 \in [0, 1]. \tag{7.9}$$

Therefore $\rho_m^2\left(d, \beta_{diff}\right)$ is equal to the sum of squared correlations between d and each component of the vector β_{diff} [Lamoš and Potocký, 1998, Chapter 7].

Let $\beta_{diff}^i, i = 1, \ldots, N = 146$ denote the difference between the level 2 principal component scores for 146 subjects and d_i are the corresponding differences in a daily measure. Then

$$\widehat{\rho_m}^2\left(d, \beta_{diff}\right) = \sum_{k=1}^{P_2} \widehat{\rho}^2\left(d, \beta_{diff_k}\right), \text{ where}$$

$$\widehat{\rho}\left(d, \beta_{diff_k}\right) = \frac{\sum_{i=1}^{N} \left(d_i - \overline{d}\right)\left(\beta_{diff_k}^i - \overline{\beta_{diff_k}}\right)}{\sqrt{\left(\sum_{i=1}^{N} \left(d_i - \overline{d}\right)\right)\left(\sum_{i=1}^{N} \left(\beta_{diff_k}^i - \overline{\beta_{diff_k}}\right)\right)}}$$

$$\overline{\beta_{diff_k}} = \frac{1}{N} \sum_{i=1}^{N} \beta_{diff_k}^i$$

is the maximum likelihood estimator for $\rho_m^2\left(d, \beta_{diff}\right)$ [Lamoš and Potocký, 1998].

Let's assume that the merged random vectors $\begin{pmatrix} \beta_{diff}^i \\ d_i \end{pmatrix}, i = 1, \ldots, N$ represent a zero–mean random sample from the $(P_2 + 1)$–dimensional normal distribution with the covariance matrix $\begin{pmatrix} 2\Lambda^{(2)} & \zeta \\ \zeta^T & \sigma^2 \end{pmatrix}$, where $\zeta \in \mathbb{R}^{P_2}, \zeta_k = \operatorname{cov}(d, \beta_{diff_k}), k = 1, \ldots, P_2$ and $\sigma^2 = \operatorname{cov}(d)$. Then the test statistics $F_T = \dfrac{N - (P_2 + 1)}{P_2} \dfrac{\widehat{\rho_m}^2\left(d, \beta_{diff}\right)}{1 - \widehat{\rho_m}^2\left(d, \beta_{diff}\right)}$ has the Fisher–Snedecor distribution with P_2 and $N-(P_2+1)$ degrees of freedom. We reject the hypothesis that $\rho_m^2\left(d, \beta_{diff}\right) = 0$ if F_T exceeds the critical value $F(\alpha, P_2, N - (P_2 + 1))$. In this case we set the critical level $\alpha = 0.01$.

However, some elements of β_{diff} may be weakly correlated with the difference in a daily measure and omitting them from (7.9) causes only minor changes in the coefficient of multiple correlation. Let's assume, that the first M level 2 functional principal components explain more than 75% of the data variability. Then $\beta_{diff_1}, \ldots, \beta_{diff_M}$ were always considered. For each β_{diff_j}, with $j > M$ we computed the second power of its correlation coefficient with d and sorted the obtained values in the descending order. Let's denote the new order by $(1), (2), (3), \ldots, (P_2 - M)$. Then we selected the first $1 \leq k \leq P_2 - M$ of the ordered elements, such that the coefficient of multiple correlation between d and $\beta^{\star}_{diff} = \left(\beta_{diff_1}, \ldots, \beta_{diff_M}, \beta_{diff_{(1)}}, \ldots, \beta_{diff_{(k)}}\right)^T$ expresses more than 95% of the second power of the original coefficient of multiple correlation $\rho^2_m\left(d, \beta_{diff}\right)$. This procedure was carried out for each combination of sleep microstate and a daily measure from Table 1.1.

sleep stage	daily measure	$\rho^2_m\left(d, \beta_{diff}\right)$
Wake	SRQ_sq	0.23
	VAS_drive	0.19
S1	VAS_aff	0.18
	VAS_drows	0.18
S2	VAS_mood	0.21
SWS	VAS_mood	0.22
	VAS_drows	0.20
	VAS_aff	0.19

Table 7.1: The second power of the coefficient of multiple correlation between the difference in the level 2 principal component scores of a sleep stage and difference in a daily measure. Only results where the coefficient was significant on level $\alpha = 0.01$ are listed.

The second power of the significant estimated coefficients of multiple correlation are listed in Tables 7.1 and 7.2. Changes in the profiles of the sleep probabilistic curves are reflected especially in changes of subjectively scored sleep quality (*SRQ_ sq*), level of mood (*VAS_ mood*), affectivity (*VAS_ aff*), drive (*VAS_ drive*) or drowsiness (*VAS_ drows*). These results follow our expectations. If a subject during the first night visits more frequently sleep microstates representing light sleep or wakefulness (due to the "first–night effect"), then it is natural to expect that his or her subjective feelings are different in contrast to the night with deeper sleep.

sleep microstate	daily measure	$\rho^2_m(d, \beta_{diff})$
1 (85% *S2*)	SRQ_sq	0.16
4 (74% *SWS*)	VAS_drows	0.19
6 (85% *Wake*)	VAS_mood	0.14
	SRQ_sq	0.12
7 (64% *S2*)	FA2	0.26
	SRQ_aq	0.22
	pul_m	0.22
	FA3	0.21
	sys_e	0.20
	dia_e	0.20
	SRQ_scom	0.18
	WB_e	0.18
	FMAT_r	0.18
9 (75% *S2*)	ACT_ts	0.23
	VAS_aff	0.20
10 (88% *S2*)	VAS_mood	0.21
	VAS_aff	0.20
	FA1	0.19

sleep microstate	daily measure	$\rho^2_m(d, \beta_{diff})$
12 (77% *S2*)	ACT_ts	0.24
	SRQ_scom	0.19
	sys_m	0.18
	VAS_drive	0.16
14 (72% *REM*)	FA2	0.21
16 (96% *SWS*)	WB_m	0.23
	VAS_mood	0.22
	VAS_aff	0.20
	WB_e	0.19
17 (98% *S2*)	WB_e	0.24
	pul_e	0.22
	VAS_aff	0.21
19 (88% *Wake*)	FA2	0.21
	SRQ_sq	0.21
	SRQ_scom	0.19
	VAS_drive	0.19
	ACT_errp	0.19
	VAS_drows	0.18
	FA1	0.17

Table 7.2: The second power of the coefficient of multiple correlation between the difference in the level 2 principal component scores of a sleep microstate and difference in a daily measure. Only results where the coefficient was significant on level $\alpha = 0.01$ are listed.

The coefficient of multiple correlation reflects how a single variable can be predicted using a linear combination of other variables. However, despite the curves alignment, the estimated correlations are at most 0.26 indicating only a moderate prediction power of differences in daily measures by differences in sleep profiles of two consecutive nights. This resembles our preliminary studies [Rošťáková et al., 2017; Rosipal et al., 2013], where we also observed that the prediction of daily measure values with the one–dimensional characteristics of the sleep structure is a difficult task.

First reason, why we are not able to successfully predict the results of a daily measure by the characteristics of the sleep structure, is the lack of a deeper information and

monitoring of the other subject's daily activities or feelings, which can affect the sleep pattern, but have only a slight influence on the monitored outcomes of the questionnaires or neuropsychological tests.

Second reason can be a high inter–subject variability or in the other words subject–based differences in sleep structure, which is not adequately taken into account. For example, existing sleep structure variation among two or more healthy sleepers may not strongly influence some of the daily measure results. On contrary, it is a strong variation of the subject's sleep pattern from his typical sleep profile which can be manifested by a strong change in his daily measure outcomes.

There are several studies showing the presence of the individual structure and patterns in sleep [Finelli et al., 2001; De Gennaro et al., 2005, 2008]. Lewandowski et al. [2013] detected a high degree of stability and individuality in the EEG spectra of two consecutive night recordings of healthy subjects from the SIESTA database. Therefore we expect that the individuality in the EEG recordings is inherited also in the sleep probabilistic curves of the PSM. However, a longer, several nights sleep monitoring would be probably needed to better capture this subject specific sleep pattern variability.

7.3 MFPCA for multilevel functional data with more general structure

In the previous section we introduced the MFPCA method for functional data with the same number of observations per subject and a strict order of observations within each subject. The method was applied to the dataset of healthy sleepers with the aim to detect whether changes in the sleep pattern between nights are mirrored in changes in daily measures. However, the results we obtained were not reasonably strong.

Functional cluster analysis and the curves synchronisation performed in Section 6.1.1 brought new insights into the existing relationships between sleep microstates and daily features in comparison to the raw data clustering. However, despite the benefit of the 2–step approach the results were not as clear as expected.

As mentioned above, the main reason for these problems is the strong individual pattern of the subjects' sleep structure. We hypothesise that an improvement in correlations between sleep structure and daily measures can be obtained by modelling a subject's specific profile present in its sleep probabilistic curves.

The MFPCA method is a candidate for the subject specific profiles extraction. How-

ever, for this purpose we need a database with a greater number of observations per subject. Large datasets with multiple observations per subject would reflect a common feature – a few observations for several subjects may be missing either due to the presence of noise or simply due to the absence of a visit of the subject. We speak about datasets where the number of observations varies between subjects (unbalanced design) or the order of observations within subjects is exchangeable (unordered visits). Therefore, we need to ask if it is possible to apply the same MFPCA algorithm also in these cases?

7.3.1 MFPCA for balanced data with unordered visits

In this section we focus on a balanced design with unordered visits. In such a case, Di et al. [2009] recommend to set $\eta_j = 0$ and the model (7.3) becomes one–way functional ANOVA

$$X_{ij}(t) = \mu(t) + Z_i(t) + W_{ij}(t), \qquad t \in T;\ \ i = 1, \ldots, I;\ \ j = 1, \ldots, J.$$

Now the estimators (7.7), (7.6) and (7.8) change into

$$\begin{aligned}
\widehat{R}_T(s,t) &= \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \left(X_{ij}(s) - \widehat{\mu}(s)\right) \left(X_{ij}(t) - \widehat{\mu}(t)\right), \\
\widehat{R}_B(s,t) &= \frac{1}{IJ(J-1)} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{l \neq j}^{J} \left(X_{ij}(s) - \widehat{\mu}(s)\right) \left(X_{il}(t) - \widehat{\mu}(t)\right), \\
\widehat{R}_W(s,t) &= \widehat{R}_T(s,t) - \widehat{R}_B(s,t), \ \ s,t \in T.
\end{aligned}$$

By computing the expected values of the proposed estimators

$$\begin{aligned}
\mathrm{E}\left(\widehat{R}_T(s,t)\right) &= \frac{I-1}{I} R_Z(s,t) + \frac{IJ-1}{IJ} R_W(s,t) \xrightarrow{I \to \infty} R_Z(s,t) + R_W(s,t) = R_T(s,t), \\
\mathrm{E}\left(\widehat{R}_Z(s,t)\right) &= \frac{I-1}{I} R_Z(s,t) - \frac{1}{IJ} R_W(s,t) \xrightarrow{I \to \infty} R_Z(s,t), \\
\mathrm{E}\left(\widehat{R}_W(s,t)\right) &= R_W(s,t)
\end{aligned}$$

we see, that $\widehat{R}_W$ is unbiased estimator for R_W, but $\widehat{R}_Z$ is only asymptotically unbiased estimator for R_Z. Therefore, when the number of subjects is small, the eigenfunctions estimated directly from $\widehat{R}_Z$ are not good representatives of the level 1 functional principal components.

In this case we recommend to interchange the order of estimators. Using the fact that

$$R_W(s,t) = \frac{1}{2} \mathrm{cov}\left(X_{ij}(s) - X_{ik}(s), X_{ij}(t) - X_{ik}(t)\right), \ \ s,t \in T \tag{7.10}$$

the unbiased estimator for R_W computed by the method of moments is

$$\widehat{R}_{W_2}(s,t) = \frac{1}{2}\frac{1}{IJ(J-1)}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{l\neq j}^{J}\left(X_{ij}(s) - X_{il}(s)\right)\left(X_{ij}(t) - X_{il}(t)\right). \tag{7.11}$$

Then it is not difficult to show that

$$\widehat{R}_{Z_2}(s,t) = \frac{I}{I-1}\left(\widehat{R}_T(s,t) - \frac{IJ-1}{IJ}\widehat{R}_{W_2}(s,t)\right),$$

is the unbiased estimators for R_Z.

7.3.2 MFPCA for unbalanced design with unordered visits

In the previous section we described the problems of unbiasedness of the covariance function estimators proposed in [Di et al., 2009] when the order of visits within each subject is exchangeable. When for each subject a different number of observations is at disposal and their order is exchangeable we can foresee similar difficulties in these covariance function estimators.

In this scenario of unbalanced data, the model (7.3) changes to

$$X_{ij}(t) = \mu(t) + Z_i(t) + W_{ij}(t), \quad t \in T;\ \ i = 1,\ldots,I;\ \ j = 1,\ldots,J_i > 1 \tag{7.12}$$

and the covariance function estimators proposed in [Di et al., 2009] have to be modified in the following way

$$\begin{aligned}
\widehat{R}_{T_{orig}}(s,t) &= \frac{1}{\sum_{i=1}^{I} J_i}\sum_{i=1}^{I}\sum_{j=1}^{J_i}\left(X_{ij}(s) - \widehat{\mu}(s)\right)\left(X_{ij}(t) - \widehat{\mu}(t)\right), \\
\widehat{R}_{B_{orig}}(s,t) &= \frac{1}{\sum_{i=1}^{I} J_i(J_i-1)}\sum_{i=1}^{I}\sum_{\substack{j,l=1\\ l\neq j}}^{J_i}\left(X_{ij}(s) - \widehat{\mu}(s)\right)\left(X_{il}(t) - \widehat{\mu}(t)\right), \qquad (7.13)\\
\widehat{R}_{W_{orig}}(s,t) &= \widehat{R}_{T_{orig}}(s,t) - \widehat{R}_{B_{orig}}(s,t), \qquad (7.14)
\end{aligned}$$

where $\widehat{\mu}(t) = \overline{X}_{\cdot\cdot}(t) = \dfrac{1}{\sum_{i=1}^{I} J_i}\displaystyle\sum_{i=1}^{I}\sum_{j=1}^{J_i} X_{ij}(t),\ \ t \in T$. Then the expected values of the covariance function estimators are

$$\begin{aligned}
\mathrm{E}\left(\widehat{R}_{T_{orig}}(s,t)\right) &= \left(1 - \frac{N_2}{N_1^2}\right)R_Z(s,t) + \left(1 - \frac{1}{N_1}\right)R_W(s,t), \\
\mathrm{E}\left(\widehat{R}_{B_{orig}}(s,t)\right) &= \left(1 - \frac{2}{N_1}\frac{N_3 - N_2}{N_2 - N_1} + \frac{N_2}{N_1^2}\right)R_Z(s,t) - \frac{1}{N_1}R_W(s,t), \\
\mathrm{E}\left(\widehat{R}_{W_{orig}}(s,t)\right) &= \left(\frac{2}{N_1}\frac{N_3 - N_2}{N_2 - N_1} - \frac{2N_2}{N_1^2}\right)R_Z(s,t) + R_W(s,t),
\end{aligned}$$

where $N_k = \sum_{i=1}^{I} J_i^k$, $k = 1, 2, 3$. If the number of subjects increases to infinity and the number of visits per subject is upper bounded, then these estimators are asymptotically unbiased estimators for R_T, R_Z and R_W, respectively. However, when the number of subjects is small the eigenfunctions estimated from (7.13) and (7.14) may be too far from the original level 1 and level 2 functional principal components.

Motivation example

Let's consider two subjects with $J_1 = 2$ and $J_2 = 10$ observations defined by the formulas

$$X_{1j}(t) = e^t + \sin(2\pi t) + \left(1 - \frac{j}{2}\right) t^2, \qquad j = 1, 2,$$
$$X_{2k}(t) = e^t + \cos(2\pi t) + 0.15\frac{k}{2}, \qquad k = 1, \ldots, 10; \quad t \in [0, 1].$$

After applying the MFPCA algorithm with the covariance function estimators (7.13) and (7.14) we estimated the subject–specific profiles $\widehat{Z}_i(t) = \widehat{\mu}(t) + \sum_{k=1}^{P_1} \widehat{\alpha}_{ik}\widehat{\phi}_k^{(1)}(t)$. Due to the fact, that the estimated principal component scores for the first level 1 principal component for both subjects are close to zero and the remaining level 1 principal components express noise only, the estimated subject–specific profiles are similar to the overall mean. Consequently, they significantly differ from the original observations (Figure 7.3). This is especially true for the first subject. Therefore it is necessary to adapt the estimators used in MFPCA for unbalanced data.

Estimator for μ

Similarly to the standard ANOVA model for unbalanced design, two alternatives for the estimator of the overall mean come into question – the weighted mean and the unweighted mean.

The weighted mean

$$\widehat{\mu}_w(t) = \frac{1}{\sum_{i=1}^{I} J_i} \sum_{i=1}^{I} \sum_{j=1}^{J_i} X_{ij}(t) = \frac{1}{\sum_{i=1}^{I} J_i} \sum_{i=1}^{I} J_i \overline{X}_{i\cdot}(t)$$

is heavily influenced by the profile of the subjects with many observations. Therefore we prefer to use the unweighted mean

$$\widehat{\mu}_{uw}(t) = \frac{1}{I} \sum_{i=1}^{I} \frac{1}{J_i} \sum_{j=1}^{J_i} X_{ij}(t) = \frac{1}{I} \sum_{i=1}^{I} \overline{X}_{i\cdot}(t) \tag{7.15}$$

which puts equal weights to all subjects regardless of their sample sizes.

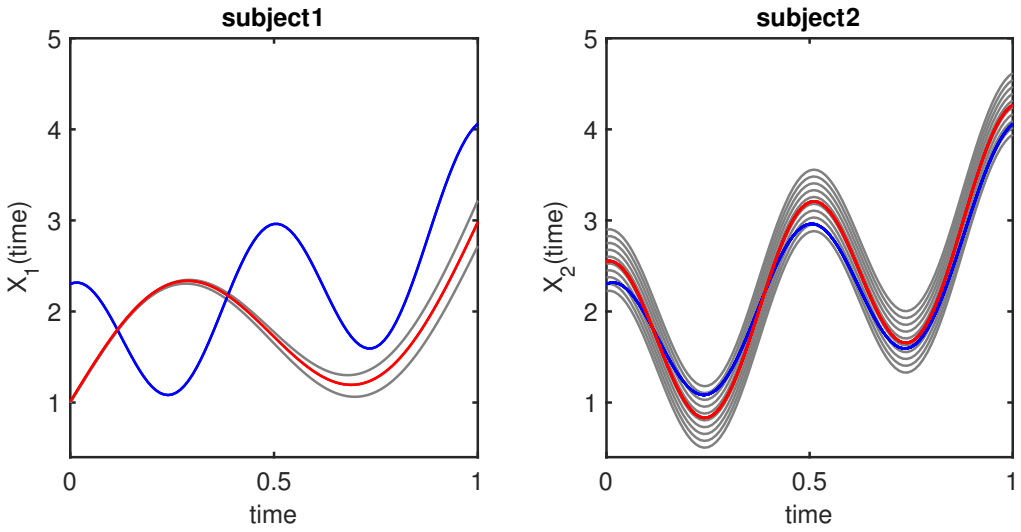


Figure 7.3: Simulated data of two subjects with two and 10 observations respectively (grey curves). The subject–specific profiles estimated by the MFPCA algorithm with the covariance function estimators (7.13) and (7.14) are depicted in blue. The red curves represent the subject–specific profiles estimated by the below suggested modifications of the covariance function estimators.

Estimator for R_W

Interchanging the order of the covariance function estimators as in the case of the balanced design with unordered visits seems to be a good idea. The estimator of (7.10) in the case of unbalanced design has the form

$$R_{W_{mm}}(s,t) = \frac{1}{2}\frac{1}{\sum_{i=1}^{I} J_i(J_i-1)} \sum_{i=1}^{I} \sum_{\substack{j,l=1 \\ l\neq j}}^{J_i} \left(X_{ij}(s) - X_{il}(s)\right)\left(X_{ij}(t) - X_{il}(t)\right) \tag{7.16}$$

and it is the unbiased estimator for R_W.

However, we were interested whether a better estimator may exist. Our goal was to find an estimator $\widehat{R}_W$ which minimises

$$\mathrm{E}\left(\|\widehat{R}_W - R_W\|^2\right) = \mathrm{E}\left(\int_T \int_T \left(\widehat{R}_W(s,t) - R_W(s,t)\right)^2 ds\, dt\right) \tag{7.17}$$

and has the form

$$\widehat{R}_W(s,t) = \frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J_i}\sum_{l\neq j}^{J_i} w_i \left(X_{ij}(s) - X_{il}(s)\right)\left(X_{ij}(t) - X_{il}(t)\right), \tag{7.18}$$

where $w_i \geq 0,\ \ i = 1, \ldots, I$ satisfy

$$\sum_{i=1}^{I} w_i J_i (J_i - 1) = 1 \tag{7.19}$$

to guarantee unbiasedness of the estimator.

Using the Fubini's theorem [Fubini, 1907] it is possible to rearrange the order of integrals and (7.17) becomes

$$\begin{aligned} \mathrm{E}\left(\|\widehat{R}_W - R_W\|^2\right) &= \int_T \int_T \mathrm{E}\left(\widehat{R}_W(s,t) - R_W(s,t)\right)^2 \, dt \, ds = \\ &= \frac{1}{4} \sum_{i,k=1}^{I} \sum_{\substack{j,l=1 \\ l \neq j}}^{J_i} \sum_{\substack{m,n=1 \\ n \neq m}}^{J_k} w_i w_k \mathrm{E}\left(D_{ijl}(s) D_{ijl}(t) D_{kmn}(s) D_{kmn}(t)\right) - R_W(s,t)^2, \end{aligned} \tag{7.20}$$

where

$$D_{ijl}(t) = X_{ij}(t) - X_{il}(t) = W_{ij}(t) - W_{il}(t), \quad t \in T; \quad i = 1, \ldots, I; \\ j, l = 1, \ldots, J_i; \ \ j \neq l.$$

However, without additional information about W_{ij} it is impossible to express (7.20) only in terms of R_W. Therefore, in the next we assume that $W_{ij}, i = 1, \ldots, I; \ \ j = 1, \ldots, J_i$ are zero–mean gaussian processes with the covariance function R_W. This implies that for each finite set of time indices $t_1, \ldots, t_k \in T,\ k \in \mathbb{N}$, the random vector $(W_{ij}(t_1), \ldots, W_{ij}(t_k))^T$ follows the multivariate normal distribution. Consequently,

$$\forall\, s, t \in T : \begin{pmatrix} D_{ijl}(s) \\ D_{ijl}(t) \\ D_{kmn}(s) \\ D_{kmn}(t) \end{pmatrix} = \begin{pmatrix} W_{ij}(s) - W_{il}(s) \\ W_{ij}(t) - W_{il}(t) \\ W_{km}(s) - W_{kn}(s) \\ W_{km}(t) - W_{kn}(t) \end{pmatrix} \sim \mathcal{N}_4\left(0, \Sigma\right),$$

where the structure of Σ depends on the choice of the indices i, j, l, k, m, n. For example if $i = k, j = m, l = n$

$$\Sigma = 2 \begin{pmatrix} R_W(s,s) & R_W(s,t) & R_W(s,s) & R_W(s,t) \\ R_W(t,s) & R_W(t,t) & R_W(t,s) & R_W(t,t) \\ R_W(s,s) & R_W(s,t) & R_W(s,s) & R_W(s,t) \\ R_W(t,s) & R_W(t,t) & R_W(t,s) & R_W(t,t) \end{pmatrix}$$

or if $i \neq k$

$$\Sigma = 2 \begin{pmatrix} R_W(s,s) & R_W(s,t) & 0 & 0 \\ R_W(t,s) & R_W(t,t) & 0 & 0 \\ 0 & 0 & R_W(s,s) & R_W(s,t) \\ 0 & 0 & R_W(t,s) & R_W(t,t) \end{pmatrix}.$$

With the aim to express the higher order moments of the multivariate normal random distribution, we apply the Isserlis' theorem [Isserlis, 1918] to (7.20) and obtain

$$\mathrm{E}\left(\|\widehat{R}_W - R_W\|^2\right) = (A+B)\sum_{i=1}^{I} w_i^2 J_i^2(J_i-1) + \frac{(N_2-N_1)(N_2-N_1-2)+1}{4}A, \quad (7.21)$$

$$A = \int_T \int_T R_W(s,t)^2 dtds,$$

$$B = \left(\int_T R_W(s,s)ds\right)^2.$$

Thanks to the assumption that $R_W : T \times T \to \mathbb{R}$ is continuous and the time interval T is closed, by using the extreme value theorem it can be shown that the integrals A and B exist and are finite.

The optimal weights $w_1, \ldots, w_I$ which minimise (7.21) subject to (7.19) can be found by the method of the Lagrange multipliers. Let us denote

$$F(w_1, \ldots, w_I, \lambda) = (A+B)\sum_{i=1}^{I} w_i^2 J_i^2(J_i-1) + \frac{(N_2-N_1)(N_2-N_1-2)+1}{4}A -$$
$$- \lambda\left(\sum_{i=1}^{I} w_i J_i(J_i-1) - 1\right),$$

where λ is a Lagrange multiplier. Setting the first derivatives of F according to $w_1, \ldots, w_I$ and λ to zero we found that

$$w_i^{opt} = \frac{1}{(N_1-I)J_i}, \quad i = 1, \ldots, I. \quad (7.22)$$

It can be proved that the matrix of the second derivatives of $F(w_1^{opt}, \ldots, w_I^{opt}, 0)$ is positive definite and therefore $\mathrm{E}\left(\|\widehat{R}_W - R_W\|^2\right)$ with subject to (7.18) and (7.19) has its global minimum in (7.22). The optimal unbiased estimator for R_W has then the form

$$\widehat{R}_{W_{opt}}(s,t) = \frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J_i}\sum_{l\neq j}^{J_i} \frac{1}{(N_1-I)J_i}\left(X_{ij}(s) - X_{il}(s)\right)\left(X_{ij}(t) - X_{il}(t)\right) \quad (7.23)$$

which can be viewed as a weighted linear combination of covariance functions within each

subject separately

$$\widehat{R}_{W_{opt}}(s,t) = \frac{1}{N_1 - I} \sum_{i=1}^{I} \sum_{j=1}^{J_i} \left(X_{ij}(s) - \overline{X}_{i\cdot}(s)\right) \left(X_{ij}(t) - \overline{X}_{i\cdot}(t)\right) = \frac{1}{N_1 - I} \sum_{i=1}^{I} J_i R^i(s,t),$$

$$\overline{X}_{i\cdot}(t) = \frac{1}{J_i} \sum_{j=1}^{J_i} X_{ij}(t),$$

$$R^i(s,t) = \frac{1}{J_i} \sum_{j=1}^{J_i} \left(X_{ij}(s) - \overline{X}_{i\cdot}(s)\right) \left(X_{ij}(t) - \overline{X}_{i\cdot}(t)\right), \quad s,t \in T.$$

Estimators for R_T and R_B

The estimator $\widehat{R}_T$ for R_T has the form

$$\widehat{R}_T(s,t) = \frac{1}{N_1} \sum_{i=1}^{I} \sum_{j=1}^{J_i} \left(X_{ij}(s) - \widehat{\mu}_{uw}(s)\right) \left(X_{ij}(t) - \widehat{\mu}_{uw}(t)\right).$$

When for each subject at least two observations are available the expected value of $\widehat{R}_T$ is

$$\mathrm{E}\left(\widehat{R}_T(s,t)\right) = \left(1 - \frac{1}{I}\right) R_B(s,t) + \left(1 - \frac{2}{N_1} + \frac{1}{I^2} \sum_{i=1}^{I} \frac{1}{J_i}\right) R_W(s,t).$$

Using the knowledge, that $\widehat{R}_{W_{opt}}$ is the unbiased estimator for R_W we can express the unbiased estimator for R_B as

$$\widehat{R}_B(s,t) = \frac{I}{I-1} \left(\widehat{R}_T(s,t) - \left(1 - \frac{2}{N_1} + \frac{1}{I^2} \sum_{i=1}^{I} \frac{1}{J_i}\right) \widehat{R}_{W_{opt}}(s,t)\right). \tag{7.24}$$

7.3.3 MFPCA for unbalanced design with unordered visits: case of a single observation for several subjects

Special case of the unbalanced design is when for some subjects only one observation is available. Then the model (7.12) becomes

$$X_{ij}(t) = \begin{cases} \mu(t) + Z_i(t) + W_{ij}(t), & \text{if } J_i > 1, \\ \mu(t) + Z_i(t), & \text{if } J_i = 1. \end{cases}$$

The estimator $\widehat{R}_{W_{opt}}$ remains the same, because its computation takes into account only subjects with more than two observations. However, small changes are needed for $\widehat{R}_B$,

because the expected value of $\widehat{R}_T$ changes in this scenario. Let L denotes the number of subjects with only one observation. Then

$$\mathrm{E}\left(\widehat{R}_T(s,t)\right) = \mathrm{E}\left(\frac{1}{N_1}\sum_{i=1}^{I}\sum_{j=1}^{J_i}\left(X_{ij}(s)-\widehat{\mu}_{uw}(s)\right)\left(X_{ij}(t)-\widehat{\mu}_{uw}(t)\right)\right) =$$

$$= \left(1-\frac{1}{I}\right)R_B(s,t) + \left(1-\frac{2}{N_1}+\frac{1}{I^2}\sum_{i=1}^{I}\frac{1}{J_i}-\frac{L}{N_1}\frac{I-2}{I}\right)R_W(s,t).$$

The unbiased estimator for $\widehat{R}_B$ is then

$$\widehat{R}_B(s,t) = \frac{I}{I-1}\left(\widehat{R}_T(s,t) - \left(1-\frac{2}{N_1}+\frac{1}{I^2}\sum_{i=1}^{I}\frac{1}{J_i}-\frac{L}{N_1}\frac{I-2}{I}\right)\widehat{R}_{W_{opt}}(s,t)\right). \quad (7.25)$$

We would like to highlight, that if for each subject more than one observation is available ($L = 0$), then (7.24) equals (7.25).

7.3.4 Application to the sleep dataset

At the beginning of this section we discussed the need for modifications in the MFPCA algorithm when used to unbalanced data and with the aim to extract the subject specific–profiles. However, the two sleep datasets introduced in Section 1.2 include only one (patients after ischemic stroke) or two observations for each subject (SIESTA database). Therefore, we don't posses an appropriate sleep database where the number of subjects' visits would be greater than two.

However, in order to demonstrate the functionality of our modified MFPCA for unbalanced data we considered the results of the cluster analysis produced by the 2–step approach. Now, each cluster represents a "subject" and curves assigned into a cluster represent repeated observation for the "subject".

The sleep probabilistic curves for Microstates 16 and 8 clustered by the 2DTW–SMTW approach and corresponding cluster–specific profiles estimated by the MFPCA method and covariance function estimators proposed in Sections 7.3.2 and 7.3.3 are depicted in Figures 7.4 and 7.5. We would like to highlight, that the outlier profile of Microstate 16 (cluster 8 in Figure 7.4) was estimated by the MFPCA method only with a negligible error.

The cluster–specific profiles estimated by the modified MFPCA method provide an alternative way for extracting cluster representatives. In comparison to the point–wise mean estimates, they can by expressed in both, the functional $\widehat{Z}_i(t) = \widehat{\mu}(t) + \sum_{k=1}^{P_1}\widehat{\alpha}_{ik}\widehat{\phi}_k^{(1)}(t)$

and vector $\widehat{\alpha}_i = (\widehat{\alpha}_{i1}, \ldots, \widehat{\alpha}_{iP_1})^T$ forms. However, the curves in a cluster are not represented with a common one–dimensional characteristics, therefore in the light of the previously used sleep to daily measures relationship investigation, we are not able to fully validate the benefit of the cluster–specific profiles extracted by the MFPCA method in comparison to the point–wise mean.

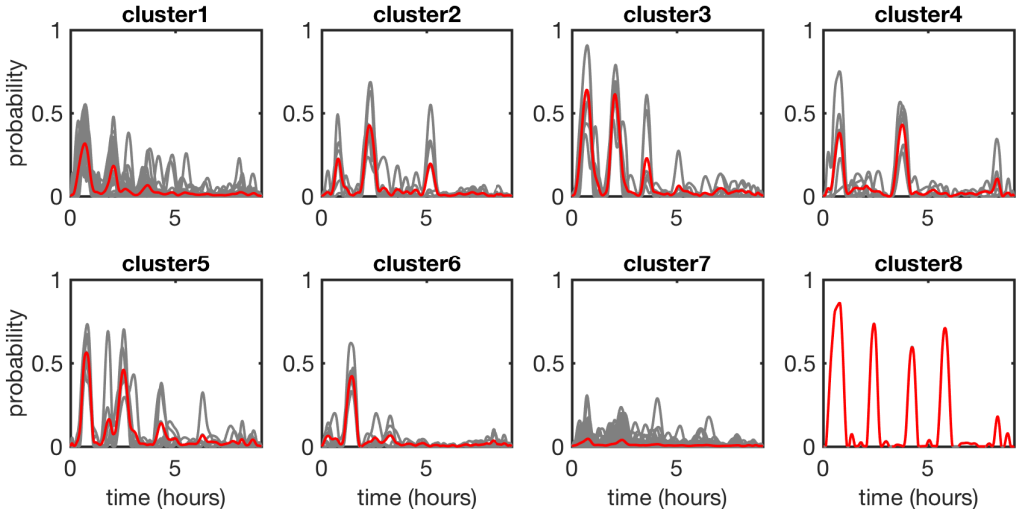


Figure 7.4: Microstate 16. Cluster analysis of 146 sleep probabilistic curves (grey) assigned into 8 clusters by the 2–step approach with the modified SMTW algorithm in the registration step and k–medoids in the clustering step (2DTW–SMTW). The cluster–specific profiles were estimated by the modified MFPCA algorithm (red curves).

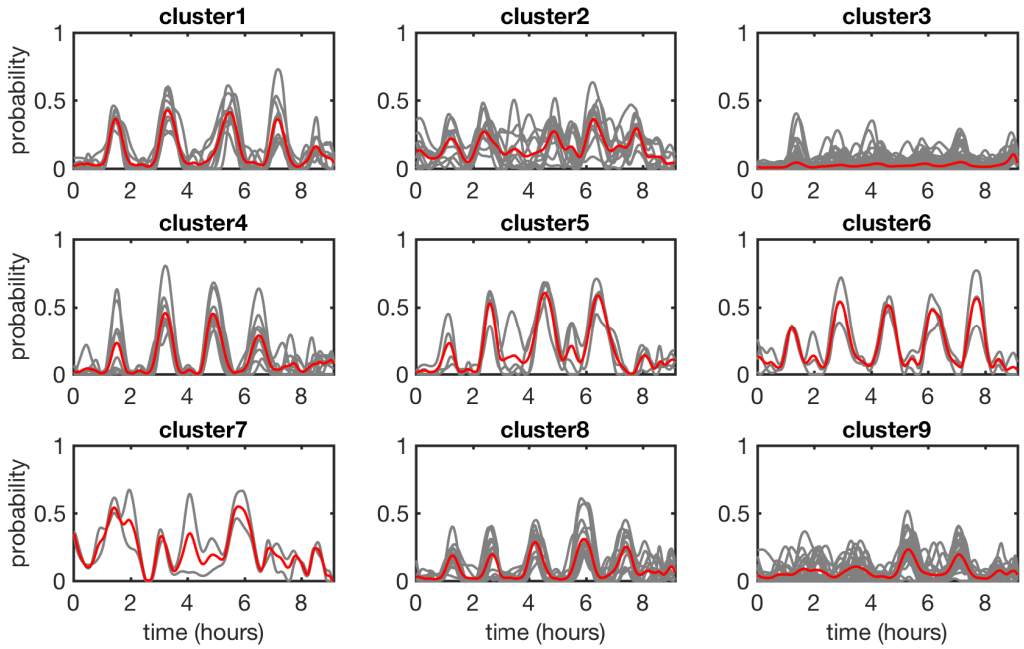


Figure 7.5: Microstate 8. Cluster analysis of 146 sleep probabilistic curves (grey) assigned into 9 clusters by the 2–step approach with the modified SMTW algorithm in the registration step and k–medoids in the clustering step (2DTW–SMTW). The cluster–specific profiles were estimated by the modified MFPCA algorithm (red curves).

Conclusion and main contributions of the thesis

Sleep as a dynamical process enters a finite number of states during night. Understanding its structure and impact on our daily life behaviour is important not only for medical practice. In this thesis we offered an alternative view on sleep structure analysis by using the methods of functional data analysis and Probabilistic sleep model [Lewandowski et al., 2012]. We would like to highlight that the analysis of the sleep structure in the functional data sense is a new approach in the area of sleep research and we are not aware of any other scientific teams which would consider this approach.

One of the major objectives of the thesis was to identify specific sleep profiles (sleep biomarkers) which significantly correlate with different physiological, demographic or daily life measures. Our previous studies [Rosipal et al., 2013; Rošťáková, 2015] showed several promising results. However, the first study did not take into account the whole dynamic of the sleep process or in [Rošťáková, 2015] we did not consider the problem of misalignment of the sleep probabilistic curves. Therefore we hypothesised, that important relationships between the sleep structure and daily measures remained hidden.

In this thesis, for a given sleep microstate we focused on finding subgroups or clusters of subjects with similar sleep probabilistic curve profiles and we tested whether there is a significant difference between results of daily measures among formed clusters. In contrast to [Rošťáková, 2015] we also analysed the impact of the curves misalignment problem on results and we considered several method to solve it.

The second approach is based on the Multilevel functional principal component analysis [Di et al., 2009] and relationship between changes in the sleep structure and changes in daily measures between two nights of a subject.

The main results and contributions of the thesis are

- We proposed our own approach for iterative combination of the curves alignment

and clustering which outperforms i) approaches where the curves alignment applied to the whole dataset precede the clustering step, and ii) approaches for simultaneous curves alignment and clustering. The benefit of our 2–step approach is also a higher flexibility of algorithmic choices in the registration step.

- The algorithm of the Self–modelling time warping method [Gervini and Gasser, 2004] does not guarantee, that the estimated warping function is strictly increasing which is in conflict with the basic assumption of the curves alignment. Therefore we considered a penalty term in the method which avoids an estimate of a non–decreasing warping function and also restricts the distance between real time and warping function.

- We showed, that the curves alignment plays important role in the analysis of the sleep structure. This is especially true for the *S2, SWS, REM* sleep stages and related sleep microstates. On the other hand, in the case of the *Wake* stage and related sleep microstates the exact occurrence of the periods of wakefulness during the night is important when detecting relationships with daily measures.

- Using the 2–step approach we detected new relationships between sleep structure and daily measures which were not observed in the case of in time misaligned sleep probabilistic curves.

- The Multilevel functional principal component analysis (MFPCA) is a method a priori developed for the detection of variability in functional data with repeated measurements. After applying the MFPCA method to the dataset of healthy sleepers we detected the "first–night sleep effect" considering several sleep microstates and all sleep stages.

- Similarly as in [Edinger et al., 2008; Rošťáková et al., 2017; Rošťáková and Rosipal, 2018] we observed that the prediction of daily measures by using characteristics of the sleep structure is a difficult task. We investigated a prediction power of the linear regression model with independent variables being the differences in the level 2 principal component scores and dependent variables being the differences in a daily measure. Then, the coefficient of multiple correlation validates the performance of the model. However, the observed correlations were at most 0.26 indicating only a moderate relationship between the changes in the sleep structure and changes in the

values of a daily measure. We hypothesise that these weak correlations are mainly influenced by the individuality in the subjects' sleep profiles.

- We adapted the MFPCA algorithm for the case when the order of observations within subjects is exchangeable or when the number of observations varies within subjects (unbalanced design). We take into account also a special case of the unbalanced design when for several subjects only one observation is available. We wrote a user–friendly MATLAB [MATLAB, 2014] script for the implementation of the modified MFPCA algorithm.

- We believe that both, the original or the modified version of the MFPCA algorithm, can be used for the detection of subject–specific profiles in sleep dataset with repeated measurements.

Finally we can conclude, that functional data analysis is a promising tool for sleep structure analysis. One of its major benefits stands from the possibility to take into the account the whole overnight sleep dynamics, which can be partially lost when considering one–dimensional sleep characteristics; being the common practice in the other existing sleep studies. To overcome and solve the discussed important problem associated with individuality of the sleep profiles, a larger database consisting of several nights sleep recordings for each subject is needed.

Resumé

Spánok je možné charakterizovať ako dynamický proces, ktorý zohráva dôležitú úlohu v našich životoch. Jeho dĺžka, kvalita a štruktúra významne ovplyvňujú naše každodenné fungovanie a v neposlednom rade aj zdravie.

Polysomnografia (PSG) je v súčasnosti najpoužívanejší diagnostický nástroj na monitorovanie a analýzu spánkového procesu. Zahŕňa súbor elektrických signálov hovoriacich o biofyziologických zmenách, ku ktorým dochádza v ľudskom organizme počas spánku. Pre potreby našej dizertačnej práce sú najdôležitejšie *elektroencefalogram* (EEG) – charakterizujúci mozgovú aktivitu – a *elektromyogram* (EMG) – reprezentujúci aktivitu svalov.

V práci sme využívali dve spánkové databázy. Prvú množinu tvorilo 146 subjektov bez výrazných spánkových porúch z databázy SIESTA [Klösch et al., 2001], ktorí strávili dve po sebe nasledujúce noci v spánkovom laboratóriu. Druhá databáza pozostávala z PSG meraní pacientov po cievnej mozgovej príhode (CMP) hospitalizovaných na I. Neurologickej klinike Lekárskej fakulty Univerzity Komenského a Univerzitnej nemocnice v Bratislave v rokoch 2013 – 2017.

V klinickej praxi sa na modelovanie spánku najčastejšie používa Rechtschaffen and Kales spánkový model (R&K) [Rechtschaffen and Kales, 1968] alebo jeho novšia verzia publikovaná Americkou asociáciou pre spánkovú medicínu (*American Academy of Sleep Medicine*, AASM) [Iber et al., 2007]. Obidva modely predpokladajú, že spánkový proces sa skladá z 5, resp. 6 spánkových stavov. Ide o stav bdelosti (*Wake* stage), ľahký spánok (*S1*, resp. *N1* v AASM), prechod k hlbokému spánku (*S2*, resp. *N2*), hlboký spánok (stavy *S3, S4*; v AASM spojené do stavu *N3*) a *REM* stav (*rapid eye movement*). Neprekrývajúce sa 30–sekundové intervaly EEG signálu a ostatných PSG signálov sú potom podľa manuálu, resp. pomocou automatického skórovacieho systému akým je napr. Somnolyzer 24x7 [Anderer et al., 2005], zaradené do jedného z uvažovaných spánkových stavov. Grafickým výstupom R&K alebo AASM je takzvaný *hypnogram*.

Obidva spánkové modely sa však vyznačujú niekoľkými nevýhodami, akými sú príliš malý počet uvažovaných spánkových stavov (5, resp. 6), diskrétna reprezentácia spánku

a s tým súvisiace skokovité prechody medzi spánkovými stavmi, heterogenita stavu *S2* a iné [Himanen and Hasan, 2000]. Z tohto dôvodu sme sa v našej práci zamerali na alternatívny prístup k modelovaniu spánku – pravdepodobnostný spánkový model (PSM) [Lewandowski et al., 2012].

Pravdepodobnostný spánkový model charakterizuje spánok pomocou pravdepodobností väčšej množiny (v našom prípade 20) spánkových stavov nazývaných aj spánkové mikrostavy. Podobne ako R&K alebo AASM, aj PSM je založený na analýze EEG, resp. EMG signálu, ale na rozdiel od klasických spánkových modelov uvažuje len intervaly dĺžky troch sekúnd. Každé z okien je následne reprezentované pomocou vektora pravdepodobností π zaradenia daného intervalu do jednotlivých mikrostavov

$$\pi = (\pi_1, \pi_2, \ldots, \pi_{20}), \quad \pi_i \in [0, 1], \quad \sum_{i=1}^{20} \pi_i = 1.$$

Okrem toho PSM umožňuje vypočítať analogické pravdepodobnosti aj pre štandardné spánkové stavy *Wake, S1, S2, SWS, REM*. Viac detailov o trénovaní a aplikácii PSM je možné nájsť v [Lewandowski et al., 2012; Rosipal et al., 2013].

Jedným z hlavných cieľov tejto práce je hľadanie špecifických spánkových profilov, tzv. spánkových biomarkerov, ktoré významne korelujú s výsledkami denných mier reprezentujúcich subjektívny, fyziologický a kognitívny stav jedinca. Pod dennými mierami rozumieme napríklad dotazník ohľadom subjetívneho hodnotenia kvality spánku [Saletu et al., 1987], nálady či nevyspatosti [Aitken, 1969], hodnoty krvného tlaku a pulzu večer pred PSG meraním a ráno po prebudení či výsledky neurokognitívnych testov zameraných na pracovnú pamäť, koncentráciu a jemnú motoriku [Grünberger, 1977; Kaufman and Lichtenberger, 2005; Greene et al., 2008].

Práce zaoberajúce sa výskumom vzťahov medzi štruktúrou spánku a dennými mierami sú zvyčajne založené na výpočte jednorozmerných spánkových charakteristík (celková doba spánku, spánková latencia, spánková eficiencia a pod.) a ich korelácií s dennými mierami [Rosipal et al., 2013; Lewandowski et al., 2012] či predikciou denných mier pomocou regresného modelu [Edinger et al., 2008; Rošťáková et al., 2017]. Iný prístup je založený na testovaní rozdielov v denných mierach medzi zhlukmi vytvorenými na základe spánkových charakteristík [Buysse et al., 2008]. Nevýhodou jednorozmerných charakteristík spánku však je, že neberú v úvahu celú časovú dynamiku spánku.

V tejto práci sme sa preto zamerali na analýzu spánku pomocou výstupov PSM. Pravdepodobnosti výskytu zvoleného spánkového mikrostavu je možné chápať ako funkciu času, pričom jej grafickú reprezentáciu nazývame pravdepodobnostná spánková krivka.

Táto spojitá reprezentácia spánku umožňuje využitie pokročilejších techník matematickej štatistiky, konkrétne funkcionálnej dátovej analýzy.

Naším cieľom bolo v rámci zvoleného spánkového mikrostavu nájsť podskupiny subjektov s podobným profilom pravdepodobnostných kriviek. Hovoríme o zhlukovej analýze kriviek. Pomocou Kruskal–Wallisovho testu sme následne testovali, či sa vytvorené zhluky signifikantne líšia aj v hodnotách zvolených denných mier.

V predchádzajúcej práci [Rošťáková, 2015] zaoberajúcej sa obdobnou analýzou štruktúry spánku sme však narazili na problém. Zvolené metódy zhlukovania často nesprávne zaradili krivky s podobným profilom do rôznych zhlukov, pretože ich dôležité charakteristiky (lokálne extrémy) boli navzájom posunuté v čase. Hovoríme, že krivky neboli časovo synchronizované.

Uvažujme dvojicu kriviek $X, Y : T \to \mathbb{R}$, bez ujmy na obecnosti nech $T = [0, 1]$. Pod časovou synchronizáciou dvojice kriviek budeme rozumieť hľadanie rastúcej synchronizačnej funkcie $h : T \to T$, ktorá minimalizuje zvolené kritérium vzdialenosti medzi krivkami, napríklad

$$h \in \operatorname{argmin}_{h^\star} \int_T \left(X(t) - (Y \circ h^\star)(t)\right)^2 dt, \quad h(0) = 0, \;\; h(1) = 1.$$

Iné príklady vzdialeností medzi krivkami je možné nájsť v [Montero and Vilar, 2014]. V práci sme sa zamerali na tri metódy navrhnuté na riešenie problému synchronizácie kriviek – *Self–modelling time warping* (SMTW) [Gervini and Gasser, 2004], metódu *Pairwise curve synchronisation* (PCS) [Müller and Tang, 2008] a metódu *Elastic time warping* (ETW) [Tucker et al., 2013].

V prípade prvej metódy sme navrhli modifikáciu algoritmu s cieľom dosiahnuť, aby odhadnutá synchronizačná funkcia (*warping function*) bola striktne rastúca. Synchronizačné funkcie odhadnuté pôvodným algoritmom SMTW boli v niektorých prípadoch len neklesajúce, t.j. obsahovali konštantné segmenty a tým pádom viedli k "natiahnutiu" určitých častí kriviek. Zároveň tým bol porušený základný predpoklad synchronizácie kriviek. Okrem toho sme uvažovali aj penalizáciu na vzdialenosť medzi reálnym časom a synchronizačnou funkciou s cieľom zabrániť synchronizácií časovo príliš vzdialených úsekov. Podobná penalizácia je zakomponovaná aj v metódach PCS [Müller and Tang, 2008] a ETW [Tucker et al., 2013].

V prvom kroku sme aplikovali vybrané metódy časovej synchronizácie kriviek na celú množinu spánkových pravdepodobnostných kriviek zvoleneného mikrostavu a následne sme časovo zosúladené krivky zaradili do zhlukov pomocou metódy k–means [Lloyd, 1982]

s verziou L^2 normy pre krivky. Bohužiaľ, tento prístup sa neukázal ako príliš efektívny a viedol k dvom extrémnym výsledkom

1. Krivky neboli dôkladne zosynchronizované a teda zhluková analýza viedla k rovnakým záverom ako v prípade bez časovej synchronizácie. Metóda SMTW je založená na synchronizácii všetkých kriviek na tzv. cieľovú krivku. V prípade veľkého počtu rôznorodých spánkových profilov je však voľba cieľovej krivky náročná, pretože cieľová krivka nemusí obsahovať charakteristiky typické pre niektoré krivky a teda celková synchronizácia je nedostatočná. Metóda PCS synchronizuje každú dvojicu kriviek osobitne. Avšak rôznorodosť spánkových profilov v našej databáze spôsobila, že pri výpočte tzv. globálnej synchronizačnej funkcie pre každú krivku sa efekty jednotlivých časových zosúladení anulovali.

2. Synchronizácia kriviek pomocou metódy ETW bez penalizácie na vzdialenosť medzi reálnym časom a synchronizačnými funkciami zas mala za následok veľké zmeny v profiloch kriviek a často viedla k "ideálnej" synchronizácií aj kriviek s rôznymi profilmi. Uvažujúc penalizáciu na vzialenosť medzi reálnym časom a synchronizačnou funkciou, zosúladenie kriviek nebolo postačujúce.

Celkovo môžeme skonštatovať, že prístup, pri ktorom časová synchronizácia kriviek predchádza zhlukovej analýze, nie je vhodný pre spánkové pravdepodobnostné krivky.

Na základe týchto výsledkov sme sa zamerali na štúdium metód, ktoré priamo kombinujú synchronizáciu kriviek a zhlukovú analýzu. Hlavná myšlienka týchto metód je synchronizácia kriviek len na zvolenú podmnožinu najpodobnejších kriviek, resp. zvolených reprezentatívnych kriviek. Metódy *k–mean alignment for curve clustering* [Sangalli et al., 2010] a *Joint probabilistic curve clustering and alignment* [Gaffney and Smyth, 2005] pracujú iba s lineárnou transformáciou času pri synchronizácii kriviek

$$h(t) = at + b, \quad a > 0, b \in \mathbb{R}$$

a teda nie sú aplikovateľné v prípade, ak sú krivky definované na rovnakom časovom intervale a desynchronizácia je nelineárneho charakteru.

Jedna z verzií metódy PCS, konkrétne *truncated Pairwise curve synchronisation* (tPCS) [Tang and Müller, 2009], je založená na synchronizácii zvolenej krivky len na podmnožinu kriviek s podobným profilom [1]. Následné zhlukovanie prebieha klasicky, napr. pomocou

[1] V prípade klasickej PCS dochádza k synchronizácií každej dvojice kriviek.

metódy k–means. Napriek tomu, že táto metóda pracuje aj s nelineárnou transformáciou času, dosiahnuté výsledky boli takmer identické so zhlukovaním nesynchronizovaných kriviek.

Keďže žiadna z existujúcich metód kombinujúcich synchronizáciu kriviek a zhlukovú analýzu sa neukázala ako vhodná pre naše spánkové dáta, rozhodli sme sa navrhnúť vlastný prístup, ktorý sma nazvali 2–kroková iteračná metóda. Ako už napovedá jej názov, táto metóda pozostáva z dvoch krokov

- zhlukovanie kriviek metódou k–medoids a matice vzdialeností skonštruovanej pomocou metódy *Dynamic time warping* [Wang and Gasser, 1997], ktorá využíva vzájomnú synchronizáciu každej dvojice kriviek.

- synchronizácia kriviek v rámci každého zhluku osobitne, napr. pomocou SMTW [Gervini and Gasser, 2004], resp. jej nami navrhovanej modifikovanej verzie, PCS [Müller and Tang, 2008] alebo ETW [Tucker et al., 2013].

Tieto dva kroky sa opakujú dovtedy, kým sa neprekročí stanovený počet iterácií, zaradenie do zhlukov ešte podlieha zmenám, resp. kým je priemerná vzdialenosť kriviek od centroidu (priemernej krivky) ich zhluku väčšia ako zvolená konštanta.

Pri aplikácii 2–krokového iteračného prístupu na obidve spánkové databázy sme pozorovali zlepšenie synchronizácie kriviek v porovnaní s metódou tPCS. Oproti zhlukovaniu časovo nesynchronizovaných kriviek náš prístup bol schopný vytvoriť homogénnejšie a lepšie odseparované zhluky. Kvalita synchronizácie bolo meraná pomocou priemernej vzdialenosti synchronizovaných kriviek od centroidu príslušného zhluku. Druhá miera, priemerná *silhouette* [Rousseeuw, 1987], zas reprezentovala kompaktnosť a vzájomnú separáciu zhlukov. *Silhouette* charakterizuje, či bola krivka zaradená do správneho zhluku (hodnota ≈ 1), patrí do iného zhluku (hodnota ≈ -1), prípadne leží na hranici medzi dvoma zhlukmi (hodnota ≈ 0) .

Zhluky vytvorené pomocou 2–krokového prístupu sa v rámci spánkových mikrostavov blízkych stavom *S2, SWS* alebo *REM* signifikantne líšili v niekoľkých denných mierach, pričom tieto rozdiely boli štatisticky nesignifikantné v prípade k–means zhlukovania nesynchronizovaných kriviek alebo metódy tPCS.

Vzťah medzi vyššou pravdepodobnosťou pre mikrostavy blízke stavu *REM* a lepším hodnotením kvality spánku bol viditeľný len v prípade použitia 2–krokového prístupu. V našej predchádzajúcej práci [Rošťáková, 2015] bol viditeľný len vzťah medzi rastúcim vekom a úbytkom hlbokého spánku, resp. s tým súvisiacim predĺžením času stráveného v

ľahkom spánku a v stave bdelosti. Vďaka dvojkrokovému prístupu sme tiež pozorovali, že starší ľudia majú nižší počet období stavu *REM*.

Takisto sme pozorovali, že oneskorený výskyt prvej fázy hlbokého spánku vedie k zvýšenej rannej ospalosti. Vplyv zvýšenej intenzity prebudení počas noci na zhoršenú náladu alebo zvýšenú ospalovsť je všeobecne známy, avšak v niektorých mikrostavoch podobných ľahkému spánku, resp. prebudeniu ho bolo možné pozorovať len vďaka 2–krokovému prístupu.

Jedinou výnimkou bol mikrostav reprezentujúci stav plnej bdelosti počas a na konci noci. V tomto prípade sa synchronizácia ukázala byť kontraproduktívna pri hľadaní vzťahov s dennými mierami. Signifikantné rozdiely v subjektívnom hodnotení kvality spánku medzi zhlukmi nesynchronizovaných kriviek zmizli po synchronizácii alebo synchronizácia nepriniesla žiadne nové výsledky. Z tohto dôvodu usudzujeme, že presná poloha a dĺžka období bdelosti počas noci je dôležitá pri detekcii vzťahov medzi štruktúrou spánku a dennými mierami. K podobnému záveru sme dospeli aj pri analýze štruktúry spánku pacientov po cievnej mozgovej príhode (CMP).

V prípade pacientov po CMP neboli prínosy 2–krokového prístupu až také jednoznačné. Predpokladáme, že to bolo spôsobené najmä nižším počtom zhlukov a celkovo nižším počtom subjektov po CMP.

Vyššia pravdepodobnosť pre ľahký spánok a príbuzné mikrostavy viedla u pacientov po CMP k lepšiemu skóre v teste motorickej aktivity. Opačný výsledok bol pozorovaný v prípade mikrostavov na rozhraní stavov *S1* a *S2*. Pacienti s vyššou pravdepodobnosťou pre mikrostavy blízke hlbokému spánku reagovali pomalšie v teste reakčných časov, na druhej strane ale dosiahli lepšie skóre v teste pracovnej pamäte [Kaufman and Lichtenberger, 2005] a cítili sa menej vyčerpaní po ukočení série neurokognitívnych testov.

Nakoniec sme ešte zistili, že pacienti po CMP so závažnejším stupňom poškodenia mali hlbší spánok ako pacienti s miernym stupňom poškodenia.

V ďalšej časti dizertačnej práce sme sa zaoberala Viacstupňovou funkcionálnou verziou metódy hlavných komponentov (*Multilevel functional principal component analysis*, MFPCA) [Di et al., 2009]. Funkcionálna verzia metódy hlavných komponentov (*Functional principal component analysis*, FPCA) sa používa na redukciu dimenzie a odhalenie hlavných smerov variability vo funkcionálnych dátach. Jej viacstupňová verzia sa používa v prípade funkcionálnych dát s opakovanými pozorovaniami. MFPCA rozlišuje dva typy variability v dátach – variabilita medzi subjektami (variabilita na 1. stupni) a variabilita v rámci opakovaných pozorovaní subjektov (variabilita na 2. stupni). Podobne ako v FPCA, aj

v MFPCA sú hlavné smery variability odhadnuté ako vlastné funkcie vhodne zvolenej kovariančnej funkcie. Konkrétne ide o kovariančnú funkciu R_B medzi dvoma pozorovaniami v rámci toho istého subjektu (1. stupeň) a kovariančnú funkciu R_W rozdielu dvoch pozorovaní v rámci toho istého subjektu (2. stupeň) [Di et al., 2009].

MFPCA metódu sme aplikovali na spánkové pravdepodobnostné krivky v jednotlivých spánkových mikrostavoch zdravých subjektov. Pre každého subjekta sme mali k dispozícií dve krivky (z prvej a druhej noci) reprezentujúce daný mikrostav. Vďaka MFPCA sme v niektorých mikrostavoch a vo všetkých klasických spánkových stavoch našli tzv. "efekt prvej noci" [Roth et al., 2005; Tamaki et al., 2016]. Keď subjekt po prvý raz spí v neznámom prostredí, jeho spánok je ľahší, subjekt je náchylnejší na prebudenie sa alebo má celkovo problém zaspať. Pravdepodobnosť výskytu ľahkého spánku (stav *S1*), stavu bdelosti a príbuzných mikrostavov bola vyššia počas prvej noci. Pre druhú noc, keď už subjekti boli viac zvyknutí na prostredie laboratória, zas dominovali stavy *S2, SWS, REM* a im podobné spánkové mikrostavy.

Okrem toho sme vypočítali koeficient mnohonásobnej korelácie medzi rozdielom hodnôt zvolenej dennej miery subjekta a prislúchajúcim rozdielom v spánkových pravdepodobnostných krivkách. Signifikantné korelácie boli pozorované medzi väčšinou spánkových mikrostavov a subjektívnym hodnotením kvality spánku, nálady či nevyspatosti.

Na druhej strane ale musíme skonštatovať, že tieto korelácie boli pomerne nízke (pod 0.26). Koeficient mnohonásobnej korelácie charakterizuje, ako dobre je možné predikovať hodnoty jednej premennej pomocou lineárnej kombinácie súboru premenných. Na základe týchto výsledkov teda môžeme usudzovať, že predikčná schopnosť modelu, ktorý zahŕňa charakteristiky reprezentujúce rozdiely v spánkovom profile subjekta medzi dvoma nocami ako nezávislú premennú a príslušné rozdiely v dennej miere ako premennú závislú, nie je veľmi vysoká.

Predikcia hodnôt denných mier pomocou charakteristík spánkového profilu subjekta je pomerne obtiažna. Edinger et al. [2008] dosiahli najvyššiu koreláciu medzi lineárnou kombináciou spánkových charakteristík (celkový čas spánku, spánková eficiencia a pod.) a výsledkami testu reakčného času na úrovni 0.21. Rosipal et al. [2013] namiesto predikcie uvažovali len klasický koeficient korelácie medzi zvolenými dennými mierami a jednorozmernými spánkovými charakteristikami, ale aj tieto korelácie nepresahovali hodnotu 0.45. Podobne nízke korelácie medzi skutočnými a predikovanými hodnotami denných mier pomocou štrukúry spánku sme pozorovali aj v [Rošťáková et al., 2017; Rošťáková and Rosipal, 2018].

Pravdepodobným dôvodom týchto nízkych korelácií je vysoká individualita spánkových profilov subjektov. Touto problemtatikou sa zoaberali napríklad Finelli et al. [2001] a De Gennaro et al. [2005, 2008]. Lewandowski et al. [2013] zas zistili pomerne vysokú podobnosť medzi spektrami EEG signálu medzi dvoma nocami subjektov z databázy SIESTA. Je preto prirodzené očakávať, že táto podobnosť sa prejaví aj v spánkových pravdepodobnostných krivkách.

Jedným z možných riešení tohto problému je odhad typického spánkového profilu zvoleného mikrostavu pre jednotlivých subjektov a jeho následné odfiltrovanie zo spánkových kriviek. Toto je však možné len za prepokladu opakovaných pozorovaní. Okrem pozložkového priemeru jednotlivých kriviek môžeme typický spánkový profil subjekta odhadnúť aj pomocou metódy MFPCA.

Databázy s opakovanými meraniami v praxi často obsahujú chýbajúce pozorovania – či už z dôvodu chybného merania alebo absencie subjekta na meraní. Teda počet pozorovaní v rámci subjektov nie je konštantný (nevyvážený dizajn), resp. poradie pozorovaní nie je striktne dané (prípad neusporiadaných pozorovaní). Tým ale dôjde k porušeniu predpokladov metódy MFPCA.

V práci sme preto navrhli modifikácie metódy MFPCA pre prípad nevyváženého dizajnu, resp. neusporiadaných pozorovaní. Zmeny sa týkali najmä odhadov kovariančných funkcií R_B a R_W. Odhad pre R_W sme zvolili tak, aby minimalizoval strednú kvadratickú odchýlku od skutočnej hodnoty kovariančnej funkcie a bol nevychýlený. V prípade R_B sme dosiahli len nevychýlenosť odhadu.

Validácia modifikovanej verzie metódy MFPCA na spánkových dátach s cieľom odhadnúť typické spánkové profily subjektov však nebola možná. Väčšina dostupných spánkových databáz obsahuje maximálne dve pozorovania pre jednotlivých subjektov a teda neumožňujú naplno využiť potenciál metódy MFPCA.

Modifikovanú MFPCA sme preto aplikovali na výsledky zhlukovej analýzy pravdepodobnostných spánkových kriviek zvolených mikrostavov. Každý zhluk teraz predstavoval "subjekta" a krivky daného zhluku zas opakované pozorovania. Pomocou MFPCA sme odhadli reprezentantov jednotlivých zhlukov. V prípade zhlukov s jedným pozorovaním metóda MFPCA odhadla reprezentantov zhlukov so zanedbateľnou odchýlkou. Odhadnutí reprezentanti zhlukov s vyššou kardinalitou zas veľmi dobre charakterizovali typické črty kriviek daného zhluku.

V závere by sme chceli skonštatovať, že analýza štruktúry spánku metódami funkcionálnej dátovej analýzy priniesla sľubné výsledky. Navrhnuté metódy a modifikácie existu-

júcich algoritmov sa ukázali ako veľmi vhodné pri aplikácii na netriviálne spánkové dáta a preto predpokladáme ich úspešnú aplikáciu aj na iné databázy nielen z oblasti analýzy spánku.

Bibliography

R. C. B. Aitken. Measurement of feelings using visual analogue scales. *Proceedings of the Royal Society of Medicine*, 62(10):989–993, 1969.

G. Anderer, S. Gruber, M. Parapatics, T. Woertz, G. Miazhynskaia, B. Klösch, J. Saletu, M. Zeitlhofer, H. Barbanoj, S. Danker-Hopfe, B. Himanen, T. Kemp, M. Penzel, D. Grőzinger, P. Kunz, A. Rappelsberger, G. Schlögl, and G. Dorffner. An E–health solution for automatic sleep classification according to Rechtschaffen and Kales: validation study of the Somnolyzer 24x7 utilizing the SIESTA database. *Neuropsychobiology*, 51(3):115–133, 2005. doi: 10.1159/000085205.

Brain Products, GmbH. BrainVision Analyser 2, 2013.

J. D. Bronzino. *Biomedical Engineering Handbook, Second Edition*, volume 1. CRC Press LLC, 2000.

T. Brott, H. P. Adams, C. P. Olinger, J. R. Marler, W. G. Barsan, J. Biller, J. Spilker, R. Holleran, R. Eberle, and V. Hertzberg. Measurements of acute cerebral infarction: a clinical examination scale. *Stroke*, 20(7):864–870, 1989. doi: 10.1161/01.STR.20.7.864.

C. Bugli, P. Lambert, and J Bigot. Curve registration using fractional polynomials with application to electroencephalograms analysis. In *Stat Discussion Paper*, 2005.

T. M. P. Bui. Differences in sleep patterns among healthy sleepers and patients after stroke. Bachelor thesis, Comenius University in Bratislava, Faculty of mathematics, physics and informatics, 2014.

T. M. P. Bui and R. Rosipal. Differences in sleep patterns among healthy sleepers and patients after ischemic stroke. In *YBERC 2014 : Proceedings of the 6th International Young Biomedical Engineers and Researchers Conference*, pages 40–45, 2014.

D. J. Buysse, M. L. Hall, P. J. Strollo, T. W. Kamarck, J. Owens, L. Lee, S. E. Reis, and K Matthews. Relationships between the Pittsburgh Sleep Quality Index (PSQI), Epworth Sleepiness Scale (ESS), and clinical/polysomnographic measures in a community sample. *Journal of clinical sleep medicine : JCSM : official publication of the American Academy of Sleep Medicine*, 4:563–71, 2008.

L. De Gennaro, M. Ferrara, F. Vecchio, G. Curcio, and M. Bertini. An electroencephalographic fingerprint of human sleep. *NeuroImage*, 26(1):114 – 122, 2005. doi: https://doi.org/10.1016/j.neuroimage.2005.01.020.

L. De Gennaro, C. Marzano, F. Fratello, F. Moroni, M. C. Pellicciari, F. Ferlazzo, S. Costa, A. Couyoumdjian, G. Curcio, E. Sforza, A. Malafosse, L. A. Finelli, P. Pasqualetti, M. Ferrara, M. Bertini, and P. M. Rossini. The electroencephalographic fingerprint of sleep is genetically determined: A twin study. *Annals of Neurology*, 64(4):455–460, 2008. doi: 10.1002/ana.21434.

Chong-Zhi Di, C. M. Crainiceanu, B. S. Caffo, and N. M. Punjabi. Multilevel functional principal component analysis. *The Annals of Applied Statistics*, 3(1):458–488, 2009. doi: 10.1214/08-AOAS206SUPP.

J. Edinger, M. K. Means, C. E. Carney, and A. D. Krystal. Psychomotor performance deficits and their relation to prior nights' sleep among individuals with primary insomnia. *Sleep*, 31:599–607, 06 2008.

L. A. Finelli, P. Achermann, and A. A. Borbély. Individual "fingerprints" in human sleep eeg topography. *Neuropsychopharmacology*, 25:57–62, 2001. doi: 10.1016/S0893-133X(01)00320-7.

G. Fubini. Sugli integrali multipli. *Rom. Acc. L. Rend. (5)*, 16(1):608–614, 1907.

S. J. Gaffney and P. Smyth. Joint probabilistic curve clustering and alignment. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 473–480. MIT Press, 2005.

D. Gervini. Self-modeling registration. http://people.uwm.edu/gervini/software/, 2004. [Online, accessed 10-March-2015].

D. Gervini and T. Gasser. Self-modeling warping functions. *Journal of the Royal Statistical Society. Series B*, 66(4):959–971, 2004. doi: 10.1111/j.1467-9868.2004.B5582.x.

D. J. Greene, A. Barnea, K. Herzberg, A. Rassis, M. Neta, A. Raz, and E. Zaidel. Measuring attention in the hemispheres: The lateralized attention network test (lant). *Brain and Cognition*, 66(1):21–31, 2008. doi: 10.1016/j.bandc.2007.05.003.

J. Grünberger. *Psychodiagnostik des Alkoholkranken. Methodischer Beitrag zur Bestimmung der Organizität in der Psychiatrie, für Ärzte, Juristen und Sozialhelfer*. Maudrich, Wien, 1977.

S. Himanen and J. Hasan. Limitations of Rechtschaffen and Kales. *Sleep Medicine Reviews*, 4(2):149–167, 2000. doi: 10.1053/smrv.1999.0086.

C. Iber, S. Ancoli-Israel, A. L. Chesson, and S. Quan. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*, 2007.

L. Isserlis. On a formula for the product–moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12(1-2):134–139, 1918. doi: 10.2307/2331932.

J. Jacques and C. Preda. Functional data clustering: A survey. *Advances in Data Analysis and Classification*, 8(3):231–255, 2014. doi: 10.1007/s11634-013-0158-y.

G. M. James. Curve alignment by moments. *The Annals of Applied Statistics*, 1(2): 480–501, 2007. doi: 10.1214/07-AOAS127.

K. Karhunen. *Über lineare Methoden in der Wahrscheinlichkeitsrechnung*. Suomalainen Tiedeakatemia, Helsinki, 1947.

A. S. Kaufman and E. O. Lichtenberger. *Assessing adolescent and adult intelligence*. John Wiley & Sons, 2005.

L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York, 1990. doi: 10.1002/9780470316801.

G. Klösch, B. Kemp, T. Penzel, A. Schlögl, P. Rappelsberger, E. Trenker, G. Gruber, J. Zeitlhofer, B. Saletu, W.M. Herrmann, S.L. Himanen, D. Kunz, M.J. Barbanoj, J. Röschke, A. Varri, and G. Dorffner. The SIESTA project polygraphic and clinical database. *Medicine and Biology Magazine*, 20(3):51–57, 2001. doi: 10.1109/51.932725.

A. Kneip and J. O. Ramsay. Combining registration and fitting for functional models. *Journal of the American Statistical Association*, 103(483):1155–1165, 2008. doi: 10.1198/016214508000000517.

T. M. Kodinariya and P. R. Makwana. Review on determining number of cluster in k–means clustering. *International Journal of Advanced Research in Computer Science and management Studies*, 1(6):90–95, 2013.

F. Lamoš and R. Potocký. *Pravdepodobnosť a matematická štatistika. Štatistické analýzy*. Univerzita Komenského Bratislava, 1998.

D. J. Levitin, R. L. Nuzzo, B. W. Vines, and J. O. Ramsay. Introduction to functional data analysis. *Canadian Psychology*, 48(3):135–155, 2007. doi: 10.1037/cp2007014.

A. Lewandowski, R. Rosipal, and G. Dorffner. Extracting more information from EEG recordings for a better description of sleep. *Computer methods and programs in biomedicine*, 108(3):961–972, 2012. doi: 10.1016/j.cmpb.2012.05.009.

A. Lewandowski, R. Rosipal, and G. Dorffner. On the individuality of sleep EEG spectra. *Journal of Psychophysiology*, 27(3):105–112, 2013. doi: 10.1027/0269-8803/a000092.

Stuart P. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28:129–137, 1982. doi: 10.1109/TIT.1982.1056489.

M. Loeve. *Fonctions aléatoires de second ordre*. Compte Rend. Acad. Sci., 1945.

J. Malmivuo and R. Plonsey. *Bioelectromagnetism - Principles and Applications of Bioelectric and Biomagnetic Fields*. Oxford University Press, New York, 1995. doi: 10.1093/acprof:oso/9780195058239.001.0001.

MATLAB. *version 8.3.0 (R2014a)*. The MathWorks Inc., Natick, Massachusetts, 2014.

MedlinePlus - Health Information from the National Library of Medicine. Aging changes in sleep. https://medlineplus.gov/ency/article/004018.htm, 2018. [Online, accessed 15-May-2018].

J. Mercer. XVI. functions of positive and negative type, and their connection the theory of integral equations. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 209(441-458):415–446, 1909. doi: 10.1098/rsta.1909.0016.

P. Montero and J. A. Vilar. TSclust: An R package for time series clustering. *Journal of Statistical Software*, 62(1):1–43, 2014. doi: 10.18637/jss.v062.i01.

H. G. Müller. PACE: principal analysis by conditional expectation. http://www.stat.ucdavis.edu/PACE/, december 2012. [Online, accessed 10-March-2015].

H. G. Müller and R. Tang. Pairwise curve synchronisation for functional data. *Biometrika*, 95(4):875–889, 2008. doi: 10.1093/biomet/asn047.

Pacific Development and Technology, LCC. T–menstat - sample form and scoring key, 2012. URL "http://publications.neurodia.com/T-MENSTAT\%20Sample\%20Form\%20and\%20Scoring\%20Key.pdf".

T. W. Parsons. *Voice and speech processing*. McGraw Hill Series in Electrical and Computer Engineering. McGraw-Hill, 1987.

Philips Respironics. Alice 5, 2010.

J. O. Ramsay and B. W. Silverman. *Functional data analysis*. Springer Series in Statistics, New York, second edition, 2005. doi: 10.1007/b98888.

A. Rechtschaffen and A. Kales. *A Manual of Standardized Terminology Techniques and Scoring System for Sleep Stages of Human Subjects*. Bethesda, MD, U.S. Dept. of Healthy, Education and Welfare, 1968.

M. Rohleder. Analýza zmien v štruktúre spánku u pacientov po mozgovej príhode. Diploma thesis, Comenius University in Bratislava, Faculty of mathematics, physics and informatics, 2016.

R. Rosipal, A. Lewandowski, and G. Dorffner. In search of objective components for sleep quality indexing in normal sleep. *Biological Psychology*, 94(1):210–220, 2013. doi: 10.1016/j.biopsycho.2013.05.014.

T. Roth, Ch. Stubbs, and J. K. Walsh. Ramelteon (TAK-375), a selective MT1/MT2–receptor agonist, reduces latency to persistent sleep in a model of transient insomnia related to a novel sleep environment. *Sleep*, 28(3):303–307, 2005.

P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(1):53 – 65, 1987. doi: 10.1016/0377-0427(87)90125-7.

Z. Rošťáková. Pravdepodobnostné modelovanie štruktúry spánku u pacientov s ložiskovou ischémiou mozgu. Diploma thesis, Comenius University in Bratislava, Faculty of mathematics, physics and informatics, 2015.

Z. Rošťáková and R. Rosipal. Time alignment as a necessary step in the analysis of sleep probabilistic curves. *Measurement Science Review*, 18(1):1–6, 2018. doi: 10.1515/msr-2018-0001.

Z. Rošťáková, G. Dorffner, Ö. Aydemir, and R. Rosipal. Estimation of sleep quality by using microstructure profiles. In Annette ten Teije, Christian Popow, John H. Holmes, and Lucia Sacchi, editors, *Artificial Intelligence in Medicine*, pages 105–115. Springer International Publishing, 2017. doi: 10.1007/978-3-319-59758-4_12.

J. Rybár, B. Cimrová, I. Farkaš, M. Varga Doležalová, and R. Rosipal. Špecifiká kognitívneho výkonu pacientov po ložiskovej ischémii mozgu. In *Kognice a umělý život XVI : Sborník z 16. ročníku konference Kognice a Umělý život (KUZ XVI)*, pages 135–141. Praha : ČVUT v Praze, 2016.

B. Saletu, G. Kindshofer, P. Anderer, and J. Grünberger. Short-term sleep laboratory studies with cinolazepam in situational insomnia induced by traffic noise. *International Journal of Clinical Pharmacology Research*, 7(5):407–418, 1987.

L. M. Sangalli, P. Secchi, S. Vantini, and V. Vitelli. k–mean alignment for curve clustering. *Computational Statistics and Data Analysis*, 54:1219 – 1233, 2010. doi: 10.1016/j.csda.2009.12.008.

A. Srivastava and E. P. Klassen. *Functional and Shape Data Analysis*. Springer–Verlag New York, 2016. doi: 10.1007/978-1-4939-4020-2.

M. Tamaki, Ji Won Bang, T. Watanabe, and Y. Sasaki. Night watch in one brain hemisphere during sleep associated with the first–night effect in humans. *Current Biology*, 26(9):1190–1194, 2016.

R. Tang and H. G. Müller. Time-synchronized clustering of gene expression trajectories. *Biostatistics*, 10(1):32–45, 2009. doi: 10.1093/biostatistics/kxn011.

J. D. Tucker. Package fdasrvf. `https://cran.r-project.org/web/packages/fdasrvf/fdasrvf.pdf`, 2016. [Online, accessed 11-February-2016].

J. Derek Tucker, W. Wu, and A. Srivastava. Generative models for functional data using phase and amplitude separation. *Computational Statistics and Data Analysis*, 61:50–66, 2013. doi: 10.1016/j.csda.2012.12.001.

D. von Zerssen, D. Köller, and E. Rey. Die befindlichkeitsskala (b–s): Ein einfaches instrument zur objektivierung von befindlichkeitsstoerungen, inbesondere im rahmen von laengsschnittuntersuchungen. *Arzneimittelforschung (Drug Research)*, 20:915–918, 1970.

R. Škoviera, Z. Rošťáková, A. Krakovská, and R. Rosipal. Spectral and complexity characteristics of sleep eeg following ischemic stroke. In *YBERC 2014 : Proceedings of the 6th International Young Biomedical Engineers and Researchers Conference*, pages 108–114, 2014.

G. Wahba. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, 1990. doi: 10.1137/1.9781611970128.

K. Wang and T. Gasser. Alignment of curves by dynamic time warping. *The Annals of Statistics*, 25(3):1251–1276, 1997. doi: 10.1214/aos/1069362747.

F. Yao, H. G. Müller, A. J. Clifford, S. R. Dueker, J. Follett, Y. Lin, B. A. Buchholz, and J. S. Vogel. Shrinkage estimation for functional principal component scores with application to the population kinetics of plasma folate. *Biometrics*, 59(3):676–685, 2003. doi: 10.1111/1541-0420.00078.

List of author's publications

Publications:

- **Z. Rošťáková** and R. Rosipal. Time alignment as a necessary step in the analysis of sleep probabilistic curves. *Measurement Science Review*, 18(1), p. 1-6,2018.

Publications in conference proceedings:

- **Z. Rošťáková**, G. Dorffner, Ö Aydemir and R. Rosipal. Estimation of sleep quality by using microstructure profiles. In *Lecture Notes in Computer Science: 16th Conference on Artificial Intelligence in Medicine, AIME 2017*, vol. 10259 LNAI, p. 105-115, 2017. ISSN 0302-9743.

- **Z. Rošťáková** and R. Rosipal. Importance of the time alignment of the sleep probabilistic curves. In *MEASUREMENT 2017 : Proceedings of the 11th International Conference on Measurement.* Editors: J. Maňka, M. Tyšler, V. Witkovský, I. Frollo. - Bratislava, Slovakia : Institute of Measurement Science, Slovak Academy of Sciences, p. 27-30, 2017. ISBN 978-80-972629-0-7.

- **Z. Rošťáková** and R. Rosipal. Multilevel functional principal component analysis for unbalanced data. In *20th European Young Statisticians Meeting. - Uppsala, Sweden: Uppsala University*, p. 51-57, 2017.

Abstracts from scientific conferences

- B. Cimrová, **Z. Rošťáková**, M. Varga Doležalová, J. Rybár, I. Farkaš and R. Rosipal. Keď cievy zradia mozog: kvalita spánku a kognitívny výkon pacientov po náhlej cievnej mozgovej príhode. In *Srdce, mozog, cievy: od normálnej k patologickej fyziológii: zborník abstraktov, Smolenice, 4.-6. apríl 2017. - Bratislava: Ústav normálnej a patologickej fyziológie SAV*, p. 23, 2017. ISBN 978-80-971699- 7-8

- **Z. Rošťáková** and R. Rosipal. Multilevel functional principal components analysis in the case of unbalanced design and small number of subjects. In *Olomoucian Days of Applied Mathematics (ODAM 2017) : Book of Abstracts. - Olomouc, Czech Republic: Palacký University in Olomouc*, p. 52, 2017.

- **Z. Rošťáková** and R. Rosipal. A novel two-step iterative approach for clustering functional data. In *22nd International Conference on Computational Statistics (COMPSTAT 2016) and the CRoNoS Summer Course and Satellite Workshop on Functional Data Analysis (CRoNoS FDA 2016): Book of Abstracts*, p. 61, 2016.

- B. Cimrová, **Z. Rošťáková**, M. Varga Doležalová, J. Rybár, I. Farkaš and R. Rosipal. Linking the quality of sleep and cognitive performance in stroke patients. In *Journal of Sleep Research, 2016*, vol. 25, suppl. 1, p. 253-254, 2016. ISSN 0962-1105

- **Z. Rošťáková**, R. Rosipal and G. Dorffner. Continuous sleep profiles clustering with a novel two-step functional data approach. In *Journal of Sleep Research*, vol. 25, suppl. 1, p. 251, 2016. ISSN 0962-1105.

- **Z. Rošťáková** and R. Rosipal. Differences in sleep microstate curves among healthy sleepers and patients after stroke. In *Clinical Neurophysiology*, 127(3), p. 126, 2016. ISSN 1388-2457.

- **Z. Rošťáková** and R. Rosipal. Differences in sleep microstate curves among patients after stroke and healthy sleepers. In *PROBASTAT 2015 : The Seventh International Conference on Mathematical Statistics. Abstracts. - Bratislava : Institute of Measurement Science, SAS*, p. 44, 2015.

Participation in congresses and conferences:

- ELITECH 2018: the 20th Conference of Doctoral Students, May 23 2018, Bratislava, Slovakia,

- EYSM 2017: the 20th European Young Statisticians Meeting, August 14–18 2017, Uppsala, Sweden,

- AIME 2017: the 16th Conference on Artificial Intelligence in Medicine, June 21–24 2017, Vienna, Austria,

- ODAM 2017: Olomoucian Days of Applied Mathematics, May 31 – June 2 2017, Olomouc, Czech Republic,

- MEASUREMENT 2017: the 11th International Conference on Measurement, May 29–31, Smolenice Castle, Slovakia,
- ESRS 2016: the 23rd Congress of the European Sleep Research Society, September 13–16 2016, Bologna, Italy,
- CRoNoS FDA 2016: the CRoNoS Summer Course and Satellite Workshop on Functional Data Analysis, August 26–28 2016, Oviedo, Spain,
- COMPSTAT 2016: the 22nd International Conference on Computational Statistics, August 23–26 2016, Oviedo, Spain,
- ECCN 2015: the 15th European Congress on Clinical Neurophysiology, September 30 – October 3 2015, Brno, Czech Republic,
- PROBASTAT 2015: the 7th International Conference on Mathematical Statistics, June 29 – July 3 2015, Smolenice Castle, Slovakia.